Reporting Understandable, Useful, and Trustworthy Results of Clinical Prediction Model Studies: Insights from Biomedical Researchers

Ivan Rahmatullah

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2024

Reading Committee:

Andrea Hartzler, Chair

Barry R. Lutz

Ari Pollack

Christopher Adolph

Program Authorized to Offer Degree:

Biomedical and Health Informatics

University of Washington

Abstract

Reporting Understandable, Useful, and Trustworthy Results of
Clinical Prediction Model Studies: Insights from Biomedical Researchers

Ivan Rahmatullah

Chair of the Supervisory Committee:
Andrea Hartzler
Department of Biomedical and Health Informatics

Despite the increasing number of clinical prediction model (CPM) studies, the quality of reporting, especially for preimpact analysis studies focusing on developing and validating CPMs in research papers, remains subpar. This poor reporting quality hinders the progression of CPM studies by impeding follow-up studies, such as external validation, impact analysis studies, and systematic reviews. While the reporting guideline for these studies, TRIPOD (Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis), emphasizes transparency, biomedical researchers advocate for CPM study results to additionally embody three quality attributes: understandable, useful, and trustworthy. Yet, the extent to which biomedical researchers perceive and ensure that CPM study results meet these quality attributes, has not been explored

This dissertation aims to bridge these gaps by identifying challenges, needs, and visualization preferences among biomedical researchers to ensure CPM study results meet these

three quality attributes. Each main chapter in this dissertation addresses a specific aim. Aim 1, presented in Chapter 4, uses a mixed-method survey to explore biomedical researchers' challenges in ensuring that CPM study results meet the three quality attributes as authors and reviewers. Aim 2, detailed in Chapter 5, involves interviews with biomedical researchers to characterize their needs to ensure the three quality attributes in CPM study results. Aim 3, outlined in Chapter 6, based on interviews with biomedical researchers, identifies visualization preferences that could enhance the quality of CPM study results. The concluding Chapter 7 summarizes these findings and their contributions to biomedical informatics, which highlight a novel approach to improve the quality of CPM study results by focusing on the three quality attributes and engaging biomedical researchers beyond traditional expert panels.

Furthermore, the dissertation includes foundational chapters setting the research stage. Chapter 1 reviews relevant prior work and outlines my motivations for this study, rooted in my experiences as a primary care clinician and biomedical researcher. Chapter 2 reports on my preliminary work through a primary care provider survey about their use of clinical prediction rules. Chapter 3 describes recruitment strategies that enhanced biomedical researchers' participation in the Chapter 4 survey, utilizing PubMed records for expanded outreach.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

First and foremost, I extend my deepest gratitude to God, Allah Subhanahu wa Ta'ala, for His boundless blessings and guidance through His last messenger, Muhammad Sallallahu alaihi wa sallam, which have been pivotal throughout my life and this journey.

This dissertation is a manifestation of the collective wisdom and invaluable support provided by numerous exceptional individuals and institutions, to whom I owe immense gratitude.

At the forefront are my family members: my parents, Abah and Ibu; my wife, Nirwesthi Ardyanthi; my sisters, Humaida and Munifah; and my nieces and nephews. Additionally, I am grateful to my extended family, including my parents-in-law, brother-in-law, and sister-in-law. Their constant love, encouragement, and belief in my capabilities have been the cornerstone of my resilience and perseverance throughout this journey. Their unwavering support has been my greatest strength.

I am profoundly grateful to my dissertation chair, Dr. Andrea Hartzler, for her guiding light throughout this journey. My sincere appreciation also goes to Dr. Barry Lutz, whose 'Flu@home' project not only provided me with invaluable experience but also steered me toward my dissertation topics. My gratitude extends to my other committee members - Dr. Kari Stephens, Dr. Ari Pollack, and Dr. Christopher Adolph, for their invaluable feedback, support, and insightful suggestions that have greatly enriched my dissertation.

Special thanks to Dr. Nancy Puttkammer and Dr. Jan Flowers from DIGI at I-TECH for offering me opportunities to broaden my informatics experience through global health and informatics projects. I am also grateful to Fulbright for the scholarship that enabled me to begin pursuing my Ph.D., a crucial step in this academic journey. The Ira Kalet and Fred Wolf Endowment Fund also deserves special recognition for supporting the conduct of surveys essential

for my research. I extend my gratitude to the primary clinicians and biomedical researchers from the WWAMI region Practice and Research Network (WPRN) for their participation in my study.

I also acknowledge the Department of Biomedical and Health Informatics at the University of Washington and extend my appreciation to the Indonesian and Muslim Communities in Seattle for their companionship and support, which have made my time here enriching and memorable.

Lastly, my heartfelt thanks to Universitas Airlangga in Indonesia, my alma mater and home institution for more than 20 years. Their encouragement and support in embarking on this journey at the University of Washington, Seattle, have been fundamental to my academic growth.

DEDICATION

This dissertation is dedicated to my beloved parents, *Abah*, H. Imam Baidhowi, and *Ibu*, Hj. Seniri. Their boundless love, wisdom, and sacrifices have been the guiding lights of my life.

Chapter 1. INTRODUCTION

1.1 BACKGROUND

Clinical prediction model (CPM) research papers in peer-reviewed journals have seen a yearly uptick over the past two decades. A basic keyword search on PubMed, using terms such as "prediction model*" [Title/Abstract] OR "prediction rule*" [Title/Abstract], showcased a notable increase in the number of publications in the past two decades, as shown in Figure 1-1 (“PubMed,” n.d.).

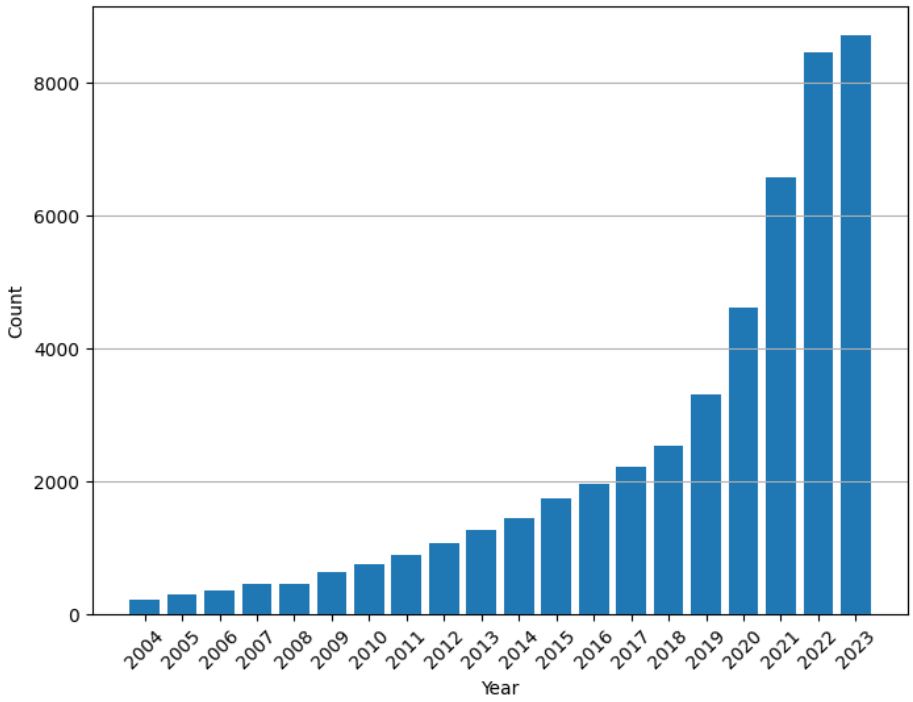

Figure 1-1 Publications for "Prediction Model*" or "Prediction Rule*" (2004-2023) in PubMed

Prediction models for clinical purposes provide risk estimates for the presence of disease (i.e., diagnosis) or an event in the future course of disease (i.e., prognosis) in individual patients (Steyerberg and Vergouwe, 2014). However, the poor quality of reporting results for CPM studies

in peer-reviewed research papers remains commonly reported by a number of systematic reviews of CPM research papers.

The systematic reviews, focusing on the preimpact analysis studies, identify the state of CPM study results reported in research papers as commonly poor and incomplete, resulting in those study reports being less 'understandable,' useful,' and 'trustworthy' (Damen et al., 2016; Dhiman et al., 2021; Heus et al., 2018; Najafabadi et al., 2020; Yang et al., 2022). Preimpact analysis studies of CPMs include studies that develop and validate the models (Cowley et al., 2019). Poor reporting quality poses challenges for the target users of CPMs, including biomedical researchers in pursuing follow-up activities such as external validation, impact analysis studies, systematic reviews, and clinicians in adopting these models in clinical practice (Dhiman et al., 2021; Maiga et al., 2019). However, the extent to which biomedical researchers perceive CPM study results and ensure that CPM study results as authors and reviewers meet these quality attributes, 'understandable,' useful,' and 'trustworthy,' has not been explored.

This dissertation focuses on investigating challenges and needs, including visualization preferences among biomedical researchers, to ensure that the results of CPM studies are understandable, useful, and trustworthy. The phrase "results in CPM studies," or "CPM study results" refers to the study results sections, including discussions and conclusions, of peer-reviewed CPM research papers and conference proceedings. Furthermore, I also refer to "preimpact analysis studies" and "impact analysis studies" as CPM studies focusing on developing/ validating CPMs and assessing the impact of CPMs in clinical practices such as studies using Randomized Control Trials (RCT) approaches, respectively.

1.2 PRIOR WORK

1.2.1 *CPMs used among clinicians*

Diagnostic errors cause up to 65,000 deaths yearly (Laposata, 2018). The Institute of Medicine (IOM) defines diagnostic error as "*the failure to (a) establish an accurate and timely explanation of the patient's health problem(s) or (b) communicate that explanation to the patient"* (Balogh et al., 2015). Unlike diagnostic errors, worsening outcomes for prognostic errors in healthcare are often undocumented and underappreciated as a failure of clinical decision-making (Graber, 2013; Khullar and Jena, 2016). A prognostic error is "*a failure to match a correctly diagnosed condition to the appropriate intervention, taking into account a patient's medical, functional, and social circumstances"* (Khullar and Jena, 2016). Improving the accuracy of diagnosis and prognosis can mitigate errors, which are among the leading causes of fatalities in medical settings (Nuñez et al., 2006; Singh et al., 2013)**.**

CPM is a type of clinical decision support (CDS) aimed at reducing diagnostic and prognostic errors during patient encounters (Medic et al., 2019). One identified factor that causes these errors is cognitive biases, which often make clinicians overly reliant on initial information and overconfident in their diagnoses, leading to diagnostic errors (Berner and Graber, 2008; Ely and Graber, 2016). CDS is a direct intervention that targets clinicians' cognitive biases of the diagnosis process during patient encounters (Karlin-Zysman et al., 2012). Education and practice feedback is an indirect intervention that addresses diagnostic errors outside patient encounters.

Of all those types of CDS, CPMs have shown improvement in diagnostic and prognostic accuracy compared to clinician judgments when comparing performance metrics and even in some trial studies (Kareemi et al., 2021; Wallace et al., 2016). CPM is a prediction model for clinical purposes to provide risk estimates for the presence of disease (i.e., diagnosis) or an event in the

future course of disease (i.e., prognosis) in individual patients (Steyerberg and Vergouwe, 2014). Clinical prediction rule (CPR) is a common term used when translating CPMs for application in clinical settings, rendering the terms interchangeable (Hemming and Taljaard, 2021), while the term CPR is more common among studies that target clinicians.

Primary care has been the target setting for many CPM studies (Keogh et al., 2014). Unfortunately, previous studies exploring the use of CPM, especially among primary care providers (PCPs) in the U.S. and globally, are outdated, being more than five years old. For example, a survey study among clinicians in the U.S. conducted in 2015 (Richardson et al., 2015), and systematic reviews exploring healthcare provider use of prediction models primarily reference studies before 2016 (Kennedy and Gallego, 2019). Thus, it is difficult to assert that they represent current usage, especially in primary care. This gap highlights the need to understand the current use of CPMs among PCPs. Addressing this gap involves assessing whether PCPs continue to use CPMs and find them beneficial. If this is the case, improving the quality of CPM study results will have a strong foundation. Such improvements in the quality of study results could ultimately enhance clinical practices, leading to better diagnostic and prognostic accuracy in primary care settings.

1.2.2 *Biomedical researchers' perspectives on the quality of the preimpact analysis study reporting*

CPM studies lie in two distinct phases: preimpact and impact analysis studies (Cowley et al., 2019). Studies in the "preimpact" phase focus on the development and internal validation of prediction models. The preimpact analysis studies focus on selecting relevant predictors, their interactions, model selections, and model performances. In contrast, studies focusing on the "impact" phase assess their practical application in real-world settings. These studies often include

RCTs or observational studies that evaluate the effects of implementing the prediction model on patient outcomes and care efficiency (Wallace et al., 2016). A study that maps the landscape of CPM studies targeting primary care found that those preimpact analyses still predominate, with 400 focusing on derivation and 200 on validation (preimpact analysis studies), and only 15 CPM studies being impact analyses (Keogh et al., 2014).

Biomedical researchers have described their expectations for the quality of reporting preimpact study results of CPM research papers in four quality attributes: transparent, understandable, useful, and trustworthy (Dhiman et al., 2021; Heus et al., 2018; Yusuf et al., 2020). These cited studies are systematic reviews that assess the qualities of individual CPM studies focusing on preimpact analyses studies against the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) criteria (Collins et al., 2015). These studies generally found that the quality of those assessed preimpact analysis studies remains poor.

Each systematic review study established expectations for quality attributes that include not only transparency, as suggested by TRIPOD, but also understandability, usefulness, and trustworthiness. Dhiman et al. noted that poor reporting quality, as indicated by incomplete adherence to the TRIPOD item list, can lead to difficulty in comprehending CPM study results, thereby highlighting 'understandable' as an expected quality attribute (Dhiman et al., 2021). Heus et al., who also observed incomplete reporting in accordance with the TRIPOD list, emphasized that poor reporting quality could hinder the use of CPM for follow-up studies and prompt the development of new models instead, thus identifying 'useful' as a quality attribute (Heus et al., 2018). Similarly, Yusuf et al., upon finding incomplete reporting as per the TRIPOD list, pointed

out that poor reporting quality could undermine trust in CPM study results, indicating 'trustworthy' as a quality attribute (Yusuf et al., 2020).

Although those systematic review studies focus on assessing the quality of CPM study reports against the TRIPOD criteria, which inherently focuses on 'transparent' attributes, each review also discusses the implications of poor reporting quality on three additional quality attributes: 'understandable,' 'useful,' and 'trustworthy.' These three quality attributes are recognized as important in previous studies evaluating information qualities, such as those assessing 'understandable,' rooted in Bloom's Taxonomies (Lord and Baviskar, 2007), 'useful' in the Technological Acceptance Model (TAM) (Larcker and Lessig, 1980), and 'trustworthy,' such as studies in Trust in Digital Information (Kelton et al., 2008).

This dissertation thoroughly investigates the challenges and needs faced by biomedical researchers to ensure that CPM study results are not only transparent but also understandable, useful, and trustworthy. By addressing these aspects, it aims to improve the quality of CPM study reporting, thereby facilitating more follow-up activities such as external validation, impact analysis studies, and systematic reviews, as well as aiding clinicians in adopting these models in clinical practice (Dhiman et al., 2021; Maiga et al., 2019).

1.2.3 *About the three quality attributes: 'understandable,' 'useful,' and 'trustworthy'*

The use of these three quality attributes of information sources -- understandable, useful, and trustworthy -- finds a more general basis in Wang and Strong's 1996 framework for data quality (Wang and Strong, 1996), which was later adapted to information quality (Ge, 2009). This framework outlines four distinct dimensions of data and information quality: intrinsic, contextual, accessibility, and representational. Each of these dimensions, when examined closely, relates to the three quality attributes—understandable, useful, and trustworthy—discussed in this

dissertation. The representational dimension closely aligns with the ‘understandable,’ contextual, and accessibility dimensions, which resonate with ‘useful,’ and intrinsic quality is closely related to ‘trustworthy. '

In this dissertation, understandable study results refer to the degree to which a reader is able to interpret, exemplify, classify, summarize, infer, compare, and explain the study results of prediction model research papers (Anderson and Krathwohl, 2001). The representational dimension, which includes aspects related to the format (concise and consistent representation) and meaning of data (interpretability and ease of understanding) (Wang and Strong, 1996), closely relates to the attribute of being understandable. This dimension emphasizes interpretability, ease of understanding, and consistent representation, aligning directly with this concept and ensuring that the data and information are presented in a way that facilitates comprehension and meaningful interpretation by the reader.

Useful study results in my dissertation are defined as the degree to which a reader is satisfied with their perceived achievement of pragmatic goals, including the goals after reading the study results of prediction model research papers (IEC, 2011). The contextual dimension, which considers the completeness, timeliness, and relevance of data within the specific context of the task at hand (Wang and Strong, 1996), closely relates to the attribute of being useful. This dimension's emphasis on assessing data quality within the context of the consumer's task aligns with the concept of usefulness, as it ensures that the data is relevant, timely, and complete for the specific goals and needs of the user.

The accessibility dimension, which emphasizes the importance of data being easily and readily accessible to data consumers (Wang and Strong, 1996), also closely relates to the attribute of being useful. Accessibility ensures that the data and information are available when needed,

which is a crucial aspect of its usefulness. If the data is not accessible, it cannot serve any practical purpose or help achieve the reader's goals, thus directly impacting its utility.

Trustworthy study results in my dissertation refer to the degree to which a reader has confidence that the study results of prediction model research papers provide information as they should (IEC, 2011). The intrinsic dimension includes accuracy, objectivity, believability, and reputation (Wang and Strong, 1996). These aspects collectively ensure that the information is inherently reliable and dependable. This closely relates to the attribute of being trustworthy. The intrinsic dimension's emphasis on accuracy, objectivity, believability, and reputation directly aligns with this definition, as these elements are fundamental to establishing trust in the information provided.

Furthermore, each attribute—understandable, useful, and trustworthy—is acknowledged as an important attribute of information quality in its own right. Each has its origins and has been studied extensively in fields that emphasize the importance of information quality. For the understandable attribute, the origin of 'understandable' can be linked to Bloom's Taxonomy of Educational Objectives, a significant work in the education field published in 1956 (Bloom, 1956). Subsequent developments expanded the use of the cognitive domain within Bloom's Taxonomy to assess students' comprehension of information throughout their learning experiences (Lord and Baviskar, 2007).

The concept of 'useful' dates back to its association with perceived usefulness in the context of management information systems, particularly in accounting, as introduced by Larcker and Lessig in the early 1980s (Larcker and Lessig, 1980). In 1985, Fred D. Davis introduced the constructs of 'ease of use' and 'perceived usefulness' as part of the Technology Acceptance Model (TAM), focusing on its application in information systems, particularly in email. (Davis, 1985).

Subsequently, various researchers have expanded the applicability of the Technology Acceptance Model (TAM) to different information technology areas and beyond. (Holden and Karsh, 2010; Venkatesh and Bala, 2008).

Regarding 'trustworthy,' the earliest scholarly work to intertwine this concept with the presentation of information was on Trust in Digital Information, presented by Kelton in 2008 (Kelton et al., 2008). Prior discussions on trust primarily centered around psychological aspects and did not incorporate information-focused perspectives (Corazzini, 1977; Gambetta, 2000). Kelton's work proposed a comprehensive model of trust in information, outlining how information sources with trustworthy characteristics such as accuracy, objectivity, validity, and consistency enhance the dependability of information for decision-making processes. Later developments in this area include assessing the trustworthiness of various sources of information, including those from information systems and libraries (Donaldson, 2016; Meeßen et al., 2020).

Frameworks that utilize the attributes 'understandable,' 'useful,' and 'trustworthy' suggest that meeting these criteria leads to a focus on action, such as decision-making or practical use. In the case of 'understandable,' frameworks like the taxonomy for evaluating user engagement in information visualization, adopting Bloom's Taxonomy, assess whether users can understand the information, as demonstrated by their ability to analyze and synthesize information from the visualization, leading to informed decision-making (Mahyar et al., 2015). Regarding 'useful,' as seen in the Technology Acceptance Model (TAM), fulfilling perceived usefulness often heightens users' attitudes toward using the system and actual usage (Davis, 1985). For 'trustworthy,' frameworks such as Trust in Digital Information and Trust in Management Information Systems (MIS) indicate that trust in information can motivate users to act on that information and enhance actual MIS use (Kelton et al., 2008; Meeßen et al., 2020).

Studies that assess the quality of information sources often focus on only one of the three quality attributes—understandable, useful, or trustworthy—rather than addressing all three simultaneously (Morgan, 2011; Porter and Donthu, 2006; Verma et al., 2018). This dissertation acknowledges the importance of all three attributes and studies them as essential qualities that should be embedded in reporting CPM study results. By integrating the quality attributes of 'understandable,' 'useful,' and 'trustworthy,' this research aims to provide a more comprehensive framework for improving the quality of CPM study reports.

1.2.4 *Reporting guidelines for the preimpact analysis studies*

TRIPOD stands as one prominent guideline to improve the quality of reporting in the preimpact analysis studies (Collins et al., 2015). Its primary objective is to enhance transparency and completeness in reporting prediction model studies. The components of the TRIPOD checklist include items ensuring the inclusion of essential information such as details about participants, variables used in the model, model-building procedures, performance measures, and visualizations for delivering results. This comprehensive approach aims to foster clarity and replicability in CPM research. Notable guidelines that fall in the same category as TRIPOD (i.e., the guideline for preimpact analysis studies) include Prediction model Risk Of Bias Assessment Tool (PROBAST) (Wolff et al., 2019) and Guidelines for Developing and Reporting Machine Learning Predictive Models in Biomedical Research (Luo et al., 2016). The other notable guidelines but not in the category of preimpact analysis studies include Critical Appraisal and Data Extraction for Systematic Reviews of Prediction Modelling Studies (CHARMS) Checklist (Moons et al., 2014), Guidelines for clinical trial protocols for interventions involving artificial intelligence (SPIRIT-AI) extension (Cruz Rivera et al., 2020).

These guidelines mostly adhere to general steps in developing health research reporting guidelines, which typically involve two main activities: conducting literature reviews and employing expert consensus through Delphi methods (Moher et al., 2010). The Delphi method is a structured process that uses a series of rounds of surveys or questionnaires to gather expert opinions and achieve consensus. Experts provide feedback, which is then synthesized and redistributed in subsequent rounds, refining the guidelines until consensus is reached. However, these steps often target specific experts recruited through professional networks and exclude a broader range of biomedical researchers who may not be accessible via Delphi methods, raising questions about whether these guidelines adequately address the actual challenges these researchers face in achieving desired qualities in their studies.

This dissertation focuses on identifying challenges and addressing the needs of biomedical researchers to ensure that CPM study results are understandable, useful, and trustworthy. By involving biomedical researchers beyond traditional expert panels, my dissertation provides a more user-centered approach to improving the quality of reporting CPM study results. The findings from this dissertation have the potential to improve reporting standards and the overall quality of CPM studies that address the identified challenges among those producing the CPM research papers, contributing to more actionable outcomes in biomedical research and facilitating more follow-up studies and implementations in clinical practices.

1.2.5 *Identifying challenges, needs, and preferences as part of Human Centered Design (HCD) principles*

HCD is an approach to system design that emphasizes understanding and focusing on the needs and requirements of end-users (FDIs, 2009). HCD consists of four design activities: Understanding and Specifying the Context of Use, Specifying User Requirements, Producing

Design Solutions, and Evaluating the Design. Identifying challenges is a key aspect of the first activity while identifying needs is central to the second.

In the "Understanding and Specifying the Context of Use" phase of HCD, identifying specific challenges faced by user groups is a key focus. This phase establishes the basis for defining user needs in subsequent stages of system design. Both quantitative methods, such as structured surveys often utilizing Likert scales (Ahuja and Jr, 2021) and qualitative methods (Muinga et al., 2021; Shrier et al., 2020), are frequently employed to identify these challenges. These techniques collectively enable a comprehensive understanding of the challenges, encompassing technical, physical, and social conditions as well as the internal context of the user (i.e., the effects, emotions, psychological needs, or intrinsic motivation) that may affect system design and use (Lallemand and Koenig, 2020; Maguire, 2001). This phase forms a robust foundation for the next phase, which involves identifying user needs.

The "Specifying User Requirements" phase of HCD focuses on understanding the tasks users will perform and the system's functional requirements (Maguire, 2001). This phase involves articulating user needs and expectations, often using natural language, diagrams, or other informal methods. These descriptions tend to be high-level, abstract, and somewhat vague, focusing on overarching goals and needs. Engaging stakeholders through observations, interviews, or questionnaires is a primary method in this phase (Faieta et al., 2024; Knowles et al., 2021). These interactions help collect vital information to shape a clear and effective design, ensuring the system aligns with its users' actual needs and expectations.

Adopting HCD principles, particularly the identification of challenges (Activity 1), needs, and preferences (Activity 2), may improve the quality of CPM study results in research papers. As suggested in Activity 1 of HCD, attempts to improve the quality of CPM study results should begin

by identifying challenges, which helps in understanding the context of use in current research papers from the perspectives of those involved in consuming and producing the CPM studies, such as authors, and reviewers. This activity should ensure the identified challenges resonate with their target users, including biomedical researchers. Once these challenges are recognized, the next logical step is to specify user requirements or needs by engaging biomedical researchers as stakeholders (Activity 2 of HCD). Once the needs are identified, as HCD suggests, producing design solutions can then be carried out. Engaging stakeholders in identifying needs and preferences ensures that the design solutions align with the challenges, which is key to the application of HCD. Producing design solutions (Activity 3 of HCD), although crucial, is beyond the scope of this dissertation.

Existing efforts to enhance the quality of reporting CPM study results, such as through guidelines like TRIPOD and PROBAST, still lack the incorporation of HCD activities. My dissertation addresses these gaps by focusing on the first two activities of the HCD approach. Therefore, this dissertation involves biomedical researchers in identifying challenges and addressing the needs and preferences to ensure the CPM study results are understandable, useful, and trustworthy.

1.2.6 *Visualizing CPMs to support understandable, useful, and trustworthy CPM study results*

The reporting of prediction models can incorporate visualizations across all six stages of their development or validation phase (e.g., data exploration, prediction selection, modeling, model exploration, model performance, and model presentations) (Lu, 2017). Data exploration involves examining the raw data to comprehend its structure, quality, and potential inconsistencies, which leads to cleaning, encoding, normalization, and transformation suitable for analysis. Feature

or predictor selection entails choosing predictors, considering various strategies, and balancing complexity with interpretability. Modeling involves selecting an appropriate algorithm, defining the model's structure, and setting parameters to fit the data. Result exploration provides a qualitative and visual understanding of how the model performs, offering insights into its applicability and limitations. Model performance emphasizes the quantitative evaluation of the developed model using statistical measures and metrics. Finally, model presentation is the concluding stage, where the model is clearly described, offering all the necessary information for others to understand, evaluate, and potentially replicate or apply the model in different contexts. Lu et al. have compiled visualizations used to present the results of prediction models, focusing on the preimpact analysis studies (Lu, 2017). However, these examples are centered on fields outside the clinical domain.

Research papers on the use of visualizations in CPMs often concentrate on specific parts of model development and validation rather than encompassing all six stages. These publications range from original research papers to reporting guideline documents. For instance, Van Belle et al. concentrate on visual designs for result exploration, employing color-based representations to illustrate the predictors' significance and their contributions to a patient's estimated risk, thereby aiding in patient management decisions (Van Belle and Van Calster, 2015). Chiang et al. focus on model presentation, particularly on interpreting seizure incident forecasts accurately (Chiang et al., 2021). On the reporting guidelines front, Bonnet et al. emphasize model presentations, suggesting visualization options for the clinical use of prediction models (Bonnett et al., 2019). Similarly, Moon et al.'s TRIPOD guidelines include various visualizations to enhance model presentation and performance reporting (Moons et al., 2015). Despite these contributions, none of

these academic works comprehensively address all six stages of the preimpact analysis studies, nor do they focus on ensuring that CPM study results are understandable, useful, and trustworthy.

Similar to studies identifying visualizations in the six stages of developing and validating prediction models, research focusing on the quality attributes of 'understandable,' 'useful,' and 'trustworthy' predominantly originates outside the clinical domain. For instance, Burns et al. utilize Bloom's taxonomy to assess the understandability of visualizations, identifying six levels of understanding: knowledge, comprehension, application, analysis, synthesis, and evaluation (Burns et al., 2020). Another study by Reeder et al. evaluates the perceived usefulness of visualizations in a public health surveillance system (Reeder et al., 2011). Additionally, Chatzimparmpas et al. explore how visualizations can enhance trust in machine-learning models (Chatzimparmpas et al., 2020). While the first and second studies concentrate on model presentations, the third study is likely focused on predictor selection, result exploration, and model presentation. To date, no studies about the use of data visualization for reporting CPM study results have been found that address all three quality attributes simultaneously or encompass all six stages of developing and validating CPMs comprehensively.

The comprehensive approach of my dissertation addresses the use of data visualizations across all six stages of CPM development and validation. This approach also aims to ensure that each stage is effectively communicated by focusing on how these visualizations address the quality attributes of being understandable, useful, and trustworthy. Furthermore, this holistic approach is intended to support follow-up activities among biomedical researchers and clinicians, such as external validation, impact analysis studies, and clinical practice applications.

1.2.7 *Munzner's Nested Model for data visualizations in CPM study results*

Munzner's Nested Model proposed a four-level guide for developing data visualizations by engaging target users (Munzner, 2009). These levels include domain problem characterization, data/operation abstraction design, encoding/interaction technique design, and algorithm design. Visualization designers may work with target users through interviews or use references to achieve these levels (Kerracher and Kennedy, 2017). The first level, domain problem characterization, involves identifying the data to be visualized and the tasks users need to perform with it. It is essential to understand the target users' needs for tasks involving visualizations and the nature, structure, and complexity of the data presentation.

The second level, data/operation abstraction design, in Munzner's Nested Model, plays a pivotal role in transforming domain-specific data into formats suitable for visualization. This level involves abstracting operations to implement tasks based on user needs identified in the first level, shaping them into visualization tasks for the study. Key to this stage is the translation of results from the development or presentation of data into effective visualizations. This step also includes defining the visualization tasks and identifying potential complexities. Through this meticulous process, the visualizations are aligned with the end-users' goals, ensuring relevance and effectiveness.

The third level, encoding/interaction technique design in Munzner's Nested Model, focuses on developing visual encoding and interaction techniques. In this stage, the key objective is to identify visual encodings that best represent the visualization tasks for the target audience. If interactivity is deemed beneficial, designing user interactions becomes crucial. The visualizations, whether interactive or not, should balance visual appeal with functionality, ensuring they are both engaging and effective for the intended users.

The fourth level, algorithm design, in Munzner's Nested Model, is crucial for creating interactive visualizations. This stage involves developing a prototype and selecting suitable algorithms that align with the visual encodings determined in the third level. Key tasks include optimizing performance for interactivity and coding the visualizations. Rigorous usability testing is essential to ensure the algorithms' correctness and efficiency, thus guaranteeing that the visualizations are user-friendly, responsive, and effective.

Applying Munzner's Nested Model to guide the presentation of visualizations in CPM research papers could significantly deviate from current practices. Currently, visualization guidance in CPM research primarily derives from literature reviews rather than methodologies that directly engage target users (Bonnett et al., 2019; Van Belle and Van Calster, 2015). Studies focusing on visualizing CPMs and involving target end users typically occur in clinical environments (Barda et al., 2020; Chiang et al., 2021). Thus, a gap exists in research about visualizing CPM results involving target users, especially biomedical researchers in academic papers, where initial communications of these studies often occur. The application of this model for reporting CPM study results highlights a potential area for future exploration, underscoring the importance of involving target users in visualizing CPM study results in research papers.

1.3 My preliminary experience with CPMs as a clinician

Coming from a background in primary care and public health, I understand the potential utility of CPMs. While I have encountered and utilized a few CPMs in clinical practice, their adoption is notably limited in primary care settings in Indonesia, where I initially practiced. CPM studies were a relatively new area when I began this dissertation project. One contributing factor could be the limited exposure to CPMs during medical training in Indonesia and a scarcity of local research on the topic. My interest grew in this area while I was involved in a project about influenza

as a research assistant when I was pursuing my Ph.D. at the University of Washington, Seattle.

1.4 My preliminary experience with CPMs as a biomedical researcher

My formative experiences as a biomedical researcher have predominantly influenced the development of this dissertation. My role as a data analyst in the "Flu@home" study led by Professor Barry Lutz in the Department of Bioengineering and Matthew Thompson in the Department of Family Medicine at the University of Washington, Seattle, from 2019 to 2021 offered a firsthand perspective on the complexities and limitations of CPMs for influenza diagnosis. This work and my contributions to two subsequent systematic reviews emphasized a prevailing focus on accuracy metrics for diagnosing influenza, often at the expense of other crucial factors like real-world applicability. These observations are particularly salient given the dual audience—biomedical researchers and clinicians—that such studies intend to reach. The following subsection is organized into three parts, each elaborating on three preliminary works in which I participated throughout my roles with the Flu@home study group.

1.4.1 *Flu@home study*

In 2020, I joined a research group at the University of Washington in Seattle that initiated a study called Flu@home (Zigman Suchsland et al., 2021). My role in this project was as a data analyst, responsible for managing, cleaning, and interpreting the data collected during the study. The study aimed to determine the accuracy of a mobile app-guided self-test using a lateral flow Rapid Diagnostic Test (RDT) for influenza, compared to a reference test involving a self-swab sample sent to a research laboratory. The motivation for conducting this study arose from the urgent need to enhance the diagnostic accuracy of home-based influenza tests. Given the significant societal burden of influenza, both in terms of health and economic impact, achieving

accurate and timely diagnosis is pivotal for the effective management and containment of the disease.

Our study employed a cross-sectional observational design. We recruited adult participants online based on their self-reported symptoms, which served as eligibility criteria for the study. The presence of a cough and at least one or more of the following symptoms—fever, chills or sweat, muscle/body aches, or feeling more tired than usual—was required for participation. The app facilitated multiple steps, including eligibility screening based on these symptoms, electronic consent, and step-by-step instructions for conducting the lateral flow RDT. Participants were also required to send a second nasal swab to our research laboratory for reference testing. Statistical analyses were conducted to evaluate the sensitivity and specificity of the RDT, and subgroup analyses were performed to explore various conditions affecting test accuracy.

Our study revealed several important findings. First, the overall accuracy of the RDT demonstrated a sensitivity of 14% and a specificity of 90%, indicating room for improvement. Second, most participants were between the ages of 25-64, predominantly female and mostly white. Third, risk factors for influenza positivity were significantly associated with the current illness interfering with daily activities. Lastly, the most reported symptoms were fatigue, cough, and runny nose, with fever and chills/sweats being significantly associated with PCR-positivity.

The study was planned to be repeated in the following year to refine its methodology and improve its diagnostic accuracy. To make the study more targeted, we conducted systematic review studies. The aim was to identify symptoms or combinations of symptoms in CPMs that could better target the eligibility criteria for study participants. This multifaceted approach was designed to enhance the study's impact by ensuring that the selected participants would provide the most valuable data for assessing the efficacy of the RDT.

1.4.2 *Systematic review and metanalysis study 1*

The second study I was involved in was a systematic review focused on the diagnostic accuracy of clinical signs, symptoms, and definitions for influenza. This study has been submitted to a journal and is waiting for reviews. In this project, I coordinated the study selection and assessment processes. This involved overseeing the initial literature search, selecting studies for full-text review, and coordinating the quality assessment using the QUADAS-2 framework (Whiting et al., 2011).

The study aimed to evaluate the diagnostic accuracy of various clinical signs, symptoms, and definitions used to diagnose influenza. It sought to provide a comprehensive understanding of how effective these clinical markers are in accurately diagnosing the condition, thereby aiding clinicians in making more informed decisions. The study was also to fill a gap in the existing literature concerning the diagnostic accuracy of clinical signs and symptoms of influenza. Given the public health implications of influenza and the need for timely and accurate diagnosis, this study aimed to consolidate existing research to provide a more definitive answer on what clinicians should look for when diagnosing influenza.

The study employed a rigorous meta-analysis methodology. An initial literature search across three major databases—PubMed, Embase, and CINAHL—yielded a total of 1,352 unique studies. After a preliminary review, 256 of these were selected for a more in-depth, full-text examination. Ultimately, 52 studies were included in the final meta-analysis, with participant sizes ranging from as few as 119 to as many as 155,866 individuals. The quality of these studies was assessed using the QUADAS-2 framework to ensure robustness and reliability.

The study undertook a comprehensive systematic review to evaluate the diagnostic accuracy of a range of signs, symptoms, and their combinations in diagnosing influenza. Our study

found that individual markers like fever and cough have limited utility in distinguishing influenza from other respiratory ailments. The most reliable individual indicators were subjective or measured fever, overall clinical impression, coryza, and fatigue. However, these markers demonstrated good sensitivity but poor specificity, making them effective in detecting but not excluding the disease. The absence of fever and cough were the most reliable signs for ruling out influenza. The study also highlighted age-dependent variations in the diagnostic accuracy of these symptoms. For example, fever was less sensitive in adults, while fatigue was more sensitive. Symptoms reliant on patient reports were generally unhelpful in infants. Among symptom combinations, cough plus fever was the most studied but showed only modest accuracy. This nuanced understanding has significant implications for clinical practice, emphasizing the need for a more comprehensive diagnostic approach. This study is still in the submission stage of publication.

1.4.3 *Systematic review and metanalysis study 2*

Our team's second systematic review shifted its focus from individual symptoms to the diagnostic accuracy of CPRs for influenza (Ebell et al., 2021). In this study, we used the term CPR instead of CPM in our publication. However, in our search strategies, we also included terms that cover both CPMs and CPRs. The study was designed to provide an updated, comprehensive review of CPRs, aiming to improve diagnostic accuracy and, consequently, patient outcomes.

Our goal was to bring the diagnostic criteria for influenza up to date, especially given the proven effectiveness of CPRs in diagnosing other medical conditions. To ensure the robustness of our findings, I coordinated a meticulous search across three major databases—PubMed, CINAHL, and EMBASE—following PRISMA guidelines. This search yielded 1,209 unique studies, from which we rigorously selected ten that met our inclusion criteria for the final analysis.

We have published results from this study that revealed a complex landscape of diagnostic accuracy for influenza (Ebell et al., 2021). We identified seven published risk scores and seven CART (Classification and Regression Tree) algorithms but found that only two had undergone any form of prospective validation. This finding highlighted a significant gap in the existing literature and underscored the need for more robust future studies. Our findings indicated that individual symptoms like fever and coryza are sensitive indicators but lack specificity. We also discovered considerable variation in diagnostic accuracy across different age groups. One of the key insights was the need for CPRs that identify three risk groups—low, moderate, and high—to better guide clinical decisions. Overall, our study serves as a cornerstone for future research, emphasizing the need for more comprehensive and validated CPRs for diagnosing influenza.

1.5 PRELIMINARY WORK

The focus of my Ph.D. dissertation on CPMs stemmed from my involvement in the Seattle Flu Study (SFS) projects. As the project had gathered data for a study enabling the development of influenza prediction models based on self-reported symptoms, I began to explore the potential of creating prediction models with this dataset. However, being an M.D. from outside the U.S., I needed to ascertain whether U.S. clinicians utilized CPMs and recognized their potential benefits before proceeding with development. This interest was intensified by literature suggesting that most CPR studies in primary care were conducted prior to 2016 (Kennedy and Gallego, 2019; Richardson et al., 2015), prompting me to investigate the current use of CPRs in primary care. Thus, I conducted a preliminary survey with primary care providers (PCPs).

This preliminary survey aims to assess the perception of use, support, and impact of CPRs among primary care providers. I conducted a survey aimed at primary care providers (PCPs) within the WWAMI region's Practice and Research Network (WPRN). This aim used the term CPR,

which is more commonly used among PCPs than CPM. Hosted on REDCap, the survey asked respondents questions focusing on the perception of use, support, and impact of CPRs in primary care. The categorization of the question includes the most useful CPRs, experiences with CPR use, the use of CPRs during patient encounters, the impact of CPR use, support for CPR usage from Electronic Health Records (EHRs), motivation, and barriers. These questions were mapped into constructs of the micro level of the Clinical Adoption Framework (CAF). This study highlights the current landscape of CPR use among PCPs, revealing key aspects of the use, support, and impact of CPRs in primary care.

Following the survey findings, I became convinced that CPRs hold significant potential to support clinical practices among PCPs. The possibility of developing influenza prediction models using SFS data appeared promising for supporting clinical practices. The survey results, revealing PCPs' main motivations for selecting CPRs to use, particularly the emphasis on ease of use alongside improved prediction accuracy, suggested that involving clinicians in the development process of CPMs could be highly beneficial. However, throughout the survey process, I learned that engaging clinicians in in-depth research activities posed feasibility challenges due to their time constraints and the limited resources at my disposal. Consequently, I began exploring alternative aspects of the survey results that could form the focus of my Ph.D. dissertation, aiming to address a pertinent issue within the scope of available resources.

As part of my survey results, I also observed a trend where selected CPRs have relatively more follow-up studies, as indicated by the number of citations from the original preimpact analysis studies that developed the models. This observation indicates that CPRs, which are subject to more follow-up pre-impact analysis studies, tend to gain wider acceptance within the biomedical research community and subsequently among clinicians, facilitated through academic publications

of these follow-up studies. As a result, clinicians recognize these CPRs as beneficial for clinical practice. Thus, to enhance the clinical recognition and adoption of more CPRs, additional follow-up studies seem to be needed, particularly for those CPRs not previously selected as the most useful in the survey. Currently, a large portion of CPM studies remains in the pre-impact analysis phase, with a limited number advancing to subsequent stages like impact analysis studies (Keogh et al., 2014). This discrepancy may be due to the poor reporting quality of the preimpact analysis studies, which might impede further follow-up studies and hinder their adoption in clinical practices (Dhiman et al., 2021; Maiga et al., 2019).

Initially, I was also inclined to engage clinicians to explore ways to improve the quality of CPM study results in research papers, potentially accelerating their adoption in clinical practice. However, I shifted my focus to biomedical researchers who highlighted poor reporting quality in the preimpact analysis studies instead of clinicians for two reasons. First, given the challenges of involving clinicians in in-depth research activities and the resources I could access to work with this population. Second, biomedical researchers hold unique roles as authors and reviewers of those CPM study results, meaning they also consume CPM study results in research papers as part of their roles. This makes them not only producers but also consumers of CPM study results. Thus, attempting to improve CPMs by involving biomedical researchers allows me to better understand the current landscape of CPM quality according to those who directly consume it, finding ways to improve the quality according to the population that also produces the CPM study results. Therefore, my dissertation examines the challenges and identifies the needs of biomedical researchers to improve the quality of reporting study results of preimpact analysis studies.

1.6 SPECIFIC AIMS

Biomedical researchers express expectations that the quality attributes for reporting the results of preimpact analysis studies should be transparent, understandable, useful, and trustworthy (Damen et al., 2016; Dhiman et al., 2021; Heus et al., 2018; Yusuf et al., 2020). While initiatives like TRIPOD have been developed to guide the reporting of CPMs in research papers, the focus is mostly on transparent quality attributes (Collins et al., 2015). However, given that researchers also suggest three quality attributes—understandable, useful, and trustworthy—in addition to transparency for reporting CPM study results, the extent to which biomedical researchers actually perceive CPM study results, as well as ensure that CPM study results meet these quality attributes as authors and reviewers, has not been explored.

My dissertation's overarching aim is to identify challenges, needs, and visualization preferences among biomedical researchers to ensure understandable, useful, and trustworthy results in CPM studies. Findings from this research will inform whether biomedical researchers encounter challenges in ensuring that CPM study results are understandable, useful, and trustworthy. If challenges exist, the study will identify the need to address these challenges and improve the quality of CPM study results. I organize my dissertation around three specific aims, each addressing distinct facets of these challenges:

1.6.1 *Aim 1. To use a survey to identify challenges among biomedical researchers to present understandable, useful, and trustworthy results of CPMs studies in research papers.*

The survey for this aim employs a mixed-method approach targeting biomedical researchers who have experience as authors or reviewers of CPM research papers in peer-reviewed biomedical journals or conference proceedings. Quantitatively, questions are divided into five categories: 1) demographic information, 2) the number of CPM papers and the years in which

participants acted as authors or peer reviewers, 3) the types of CPM studies participants have experienced as authors or peer reviewers, 4) the frequency with which participants find study results to be understandable, useful, and trustworthy in such papers, and 5) the difficulties participants may encounter as authors or reviewers to ensure understandable, useful, and trustworthy results in CPM studies. Qualitatively, open-ended questions probe the reasoning behind responses to the questions in category 5 (i.e., the difficulties participants may encounter as authors or reviewers to ensure understandable, useful, and trustworthy results in CPM studies). The findings of this aim contribute to quantitatively demonstrating significantly different challenges across three attributes—understandable, useful, and trustworthy—as key quality attributes of CPM study results among biomedical researchers. The inclusion of qualitative analysis of the open-ended survey responses adds depth to understanding the complex challenges researchers face in ensuring CPM results meet these three quality attributes.

1.6.2 *Aim 2. To use interviews to characterize the need for understandable, useful, and trustworthy study results for CPM research papers among biomedical researchers.*

Following the survey in Aim 2, respondents were invited for a 60-minute remote interview to capture the need and visualization preferences for presenting understandable, useful, and trustworthy results in CPM studies. The interview question for this aim focuses on identifying biomedical researchers' need to ensure that CPM study results are understandable, useful, and trustworthy. Data analysis included transcribing the recordings of the interviews and analyzing the transcripts using a directed content analysis, which is a deductive approach of qualitative analysis (Hsieh and Shannon, 2005). The outcome of this aim is a description of biomedical researchers' needs for understandable, useful, and trustworthy CPM study results.

1.6.3 *Aim 3. To use interviews to identify visualization preferences that support ensuring CPM study results are understandable, useful, and trustworthy.*

This specific aim utilized the same data from the interviews in Aim 2. While the deductive qualitative analysis I conducted in Aim 2 focused on describing biomedical researchers' needs to ensure understandable, useful, and trustworthy CPM study results, the analysis for Aim 3 focused on the following research questions: 1) how do biomedical researchers perceive the use of visualizations in presenting the results of CPM studies, 2) which visualization tasks and encodings are preferred by biomedical researchers in the presentation of results of CPM studies, and 3) how can visualizations preferred by biomedical researchers ensure understandable, useful, and trustworthy results in CPM studies. For this aim, I employed the directed content analysis to assess biomedical researchers' preferences in using visualizations to present CPM study results and identify preferred visualization tasks and their corresponding visual encoding that ensure CPM study results are understandable, useful, and trustworthy. The outcome of this aim is a description of the visualization tasks and their visual encoding preferred by biomedical researchers to ensure that CPM study results are understandable, useful, and trustworthy.

1.7 DISSERTATION OUTLINE

Building on my hands-on experiences that have influenced my perspective on CPMs, I structured my dissertation into seven chapters:

Chapter 1: *Introduction*. The current chapter provides an overview of prior work, my experiences that led to this dissertation, and this dissertation's objectives.

Chapter 2: *"Assessing the perception of use, support, and impact of CPR among Primary Care Providers: A preliminary study"* reports on my preliminary study that administers a survey to

PCPs to understand the current landscape of CPR use, focusing on the perception of use, support and impacts of CPRs in primary care.

Chapter 3: *"Using PubMed to bolster the recruitment of biomedical researchers,"* describes innovative ways to recruit biomedical researchers to capture their experiences in producing understandable, useful, and trustworthy results of CPM studies in research papers. I reflect on my use of PubMed in Aim 1 compared to traditional recruitment methods for survey research.

Chapter 4 (Aim 1): "Administering a mixed methods survey to understand challenges experienced by biomedical researchers in presenting understandable, useful, and trustworthy results in CPM studies" report on Aim 1, which employs a mixed-method approach to identify the inherent challenges biomedical researchers face when producing CPM papers that are understandable, useful, and trustworthy.

Chapter 5 (Aim 2): "Characterizing the needs of biomedical researchers for understandable, useful, and trustworthy results in CPM studies: An interview study" reports on qualitative themes regarding needs.

Chapter 6 (Aim 3): *"Identifying visualization preferences among biomedical researchers for understandable, useful, and trustworthy results in CPM studies: an interview study"* reports on qualitative themes regarding preferred visualizations to achieve understandable, useful, and trustworthy results in CPM studies.

Chapter 7: *Conclusion and contribution.* This chapter summarizes the findings and contributions to biomedical informatics. The contributions emphasize showcasing a novel survey recruitment approach that engaged a broader range of biomedical researchers beyond traditional expert panels and recommending the extension of TRIPOD guidelines based on the identified needs and visualization preferences among biomedical researchers to improve the

reporting quality of CPM study results across three quality attributes, understandable, useful and trustworthy.

Chapter 2. ASSESSING THE PERCEPTION OF USE, SUPPORT, AND IMPACT OF CLINICAL PREDICTION RULES AMONG PRIMARY CARE PROVIDERS: A PRELIMINARY STUDY

2.1 INTRODUCTION

Diagnostic and prognostic errors are medical errors that result in fatalities among patients (Nuñez et al., 2006; Singh et al., 2013). Diagnostic errors might cause up to 65,000 deaths each year (Laposata, 2018). In primary care, the estimate of U.S. adults affected by diagnostic error is 1 in 20 (Singh et al., 2014). A prognostic error is a failure to match a correctly diagnosed condition to the appropriate intervention, taking into account a patient's medical, functional, and social circumstances (Khullar and Jena, 2016). Unlike diagnostic errors, worsening outcomes for prognostic errors in healthcare are often undocumented and underappreciated as a failure of clinical decision-making (Graber, 2013; Khullar and Jena, 2016).

Clinicians may use CPRs, such as the CENTOR score for diagnosing streptococcal pharyngitis and the Framingham Score, to estimate cardiovascular disease risks (Plüddemann et al., 2014). CPR is a more common term than CPMs among clinicians (Hemming and Taljaard, 2021). Studies that mapped the landscape of CPR studies found that most of these studies targeted primary care settings (Cowley et al., 2019; Keogh et al., 2014). The use of CPRs to support clinicians has shown improvements in the accuracy of diagnoses and prognoses compared to clinician judgments alone in primary care settings (Wallace et al., 2016).

Clinicians generally favor the implementation of CPRs. This is supported by studies that highlight the ease of use and utility of CPRs in clinical practices (Plüddemann et al., 2014;

Richardson et al., 2015). However, despite this positive outlook, persistent barriers like cognitive burden and clinical workflow disruptions hinder widespread adoption (Kappen et al., 2016; Kennedy and Gallego, 2019). The implementation of CPRs involves complexities that go beyond clinicians' positive attitudes toward motivation and barriers. To better understand these complexities, I used the Clinical Adoption Framework (CAF). CAF offers a multi-level approach for a comprehensive analysis of the various elements influencing the adoption and effective use of eHealth applications in clinical practice (Lau and Price, 2017). According to the framework, CPRs can be treated as an application in healthcare to support decision making and integrated into eHealth applications through EHR (Solomon et al., 2023).

As shown in Figure 2.1, CAF is a multi-level model for assessing eHealth system adoption in healthcare that can be adapted for CPRs. The micro level focuses on factors related to the effective use of eHealth in clinical settings that include Health Information System (HIS) support quality (system, information, and service quality), usage quality (use pattern, usage intention, user satisfaction, ease of use, competency), and net benefits (improvements in care quality, access, productive, patient safety, and health outcomes) (O'Donnell et al., 2018; van Mens et al., 2020). The meso level deals with People (individual characteristics, expectations), Organization (strategic fit, culture), and Implementation (adoption stages, project management). Finally, the macro level includes healthcare standards, funding and incentives, legislation/policy and governance (regulatory influences), and Socioeconomic trends (public expectations, socio-political climate).

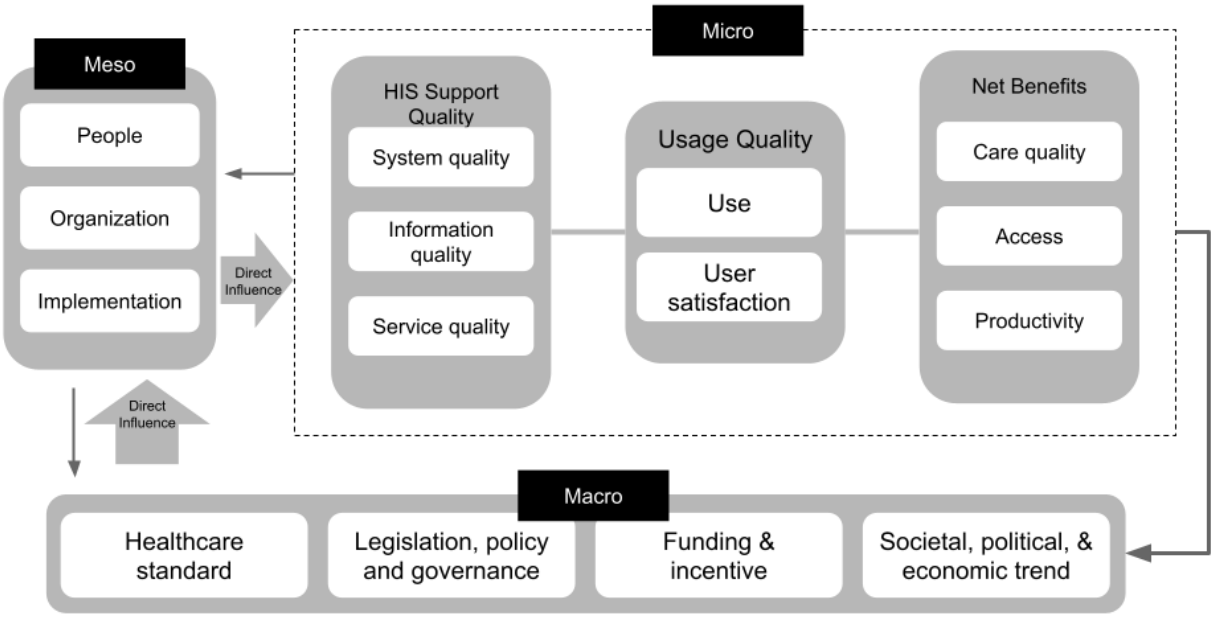

Figure 2-1 CAF Framework

Previous CPR studies targeting primary care providers (PCPs) mostly focused on identifying motivating factors and barriers to CPR use and were mostly from 2015 or earlier (Kappen et al., 2016; Kennedy and Gallego, 2019; Plüddemann et al., 2014; Richardson et al., 2015). Those studies on CPRs largely overlooked aspects beyond motivation and barriers, missing key aspects related to the effective use of CPR in clinical settings. The micro level of CAF suggests that CPR as an application in healthcare should consider the following factors: Usage Quality, HIS Support Quality, and Net Benefits, representing use, support, and impacts respectively. Identifying perceptions beyond just the motivation and barriers, and focusing on these three key aspects (i.e., perception of use, support, and impacts of CPRs), can provide a more comprehensive understanding of the current state of CPR use in primary care.

Improving preimpact analysis studies through understanding CPR use

CPMs that inform the use of CPRs in clinical practice have witnessed significant advancements, notably with the publication of guidelines for reporting CPM studies, such as TRIPOD and PROBAST (Moons et al., 2015; Wolff et al., 2019) and the integration of advanced machine learning models into CPM studies (Klement and El Emam, 2023). Despite these advancements, concerns about the poor reporting quality of CPM studies, particularly the preimpact analysis studies, persist. This issue could influence the continued use of CPRs that need to adapt to new clinical challenges, ultimately necessitating the development of new CPMs to inform CPRs for clinical practice.

Poor-quality preimpact analysis studies, if not addressed, could hinder follow-up studies, such as impact analysis studies, and jeopardize clinical practices by providing outdated or irrelevant CPRs that fail to meet current clinical challenges. Therefore, understanding the current state of CPR use among PCPs, by examining the use, support, and impacts of CPRs in primary care, may yield findings that motivate further research aimed at enhancing the quality of preimpact analysis studies.

2.2 STUDY OBJECTIVE

This study aims to assess perceptions of use, support, and impact of CPRs among PCPs. Each aspect of the assessment is informed by a construct of the micro level in the CAF, with Usage Quality guiding the perceptions of 'use,' HIS Support Quality informing the perceptions of 'support,' and Net Benefits focusing on the perceptions of 'impact.' This understanding is expected to find and confirm that CPRs are still in use and assist PCPs in clinical practices. Consequently, this finding supports the notion that studies aimed at improving the reporting qualities of CPM study results will be beneficial, as they are grounded in the reality of clinical practice where CPRs

continue to be utilized. The research question for this aim is: *How do PCPs perceive the use, support, and impact of CPRs in their clinical practice?*

2.3 METHODS

I conducted a survey targeting PCPs in the WWAMI (Washington, Wyoming, Alaska, Montana, and Idaho) region Practice and Research Network (WPRN) to assess their perceptions about the adoption of CPRs in clinical practice. WPRN serves as a coordinating hub, offering necessary tools and connections for conducting research in primary care and community-based clinical settings (ITHS, n.d.). This network spans primary care clinics and providers in the five states of the WWAMI region, covering both urban and rural locations, community health centers, private practices, academic affiliates, and government clinics. The survey's development and recruitment were done in collaboration with WPRN administrators to ensure that it met their standards, such as ensuring that the survey contained no more than ten close-ended questions and took just a few minutes to complete.

Recruitment

I recruited PCPs from WPRN without restrictions on years of practice or experience of using CPRs. The recruitment emails were sent twice through the WPRN mailing list. Each email included consent information, a study description, and a link to the online survey hosted on REDCap.

Data Collection

I collected survey data using the REDCap platform. Appendix A shows the survey instrument. In the first part of the survey, respondents answered their background characteristics (i.e., type of degree and years of practice). The second part of the instrument focused on perceptions of use, support, and benefits based on the CAF micro-level constructs HIS Support

Quality (support), Usage Quality ('use'), and Net Benefits ('impact'). These constructs are distributed into categories of questions within groups for the perception of use, support, and impact. The first part of the questions is about the Perception of Use, which includes questions about the most useful CPRs, experiences with CPR usage, and the use of CPRs. The second part addresses the impact of CPR use, which focuses on questions about the use of CPRs during patient encounters. The third focuses on perceptions of support for CPR usage from Electronic Health Records (EHRs). Previous studies in CPRs also inform the questions in the survey (Kennedy and Gallego, 2019; Richardson et al., 2015; Wallace and Johansen, 2018).

For the question about the most useful CPRs, I presented 15 validated and commonly used CPRs in U.S. primary care based on a prior survey with clinicians (Richardson et al., 2015) and references found during the development of the study, as shown in Table 2-1. Respondents selected the top three as the most useful from the list. Table 2-1 lists the 15 CPRs included in the survey and their references.

Table 2.1 CPRs included in the survey

CPR	Description
Flu Score	Support seasonal flu diagnosis using influenza-like illness, such as cough and fever (Ebell et al., 2012).
Walsh Rule/ Centor Score	Distinguish patients with viral infections and those suspected of having streptococcal pharyngitis (McGinn et al., 2003).
Bacterial Pneumonia Score	Identify children with pneumonia who need antibiotic medication (Moreno et al., 2006).
Framingham Score/ QRISK 2	Provide individualized cardiovascular disease risks to high-risk patients (Collins and Altman, 2010).
MICE Rule	Optimize referral to echocardiography for patients with suspected heart failure (Collins and Altman, 2010).
GERD Score	Support the diagnosis of Gastro-Esophageal Reflux Disease (GERD) for patients with upper gastrointestinal complaints (N. Horowitz et al., 2007).
PHQ-9/ PREDICT-NL	Detect patients with major depressive disorder (Zuithoff et al., 2009).

CPR	Description
The Breast Cancer Risk Assessment Tool (BCRAT)/ Gail model/ Nottingham Prognostic Index	Predict breast cancer risk and prognosis (Schonberg et al., 2015).
Risk Assessment Tool (RAT)/ QCancer	Estimate risk for patients with symptoms of possible cancer (Hamilton, 2009).
Diabetic Foot Screen Tool	Identify patients at risk for diabetic foot ulcers (Murphy et al., 2012).
PredictAL	Predict the onset of hazardous alcohol drinking over 12 months (Bellón et al., 2017).
Marburg Heart Score	Rules out coronary artery disease in patients with chest pain (Haasenritter et al., 2015).
CHADS2	Classification schemes that estimate stroke risk in patients with Atrial Fibrillation (Gage et al., 2001).
"CURB" severity score	Stratify patients with Community-Acquired Pneumonia CAP into different management groups using a six-point score based on confusion, urea, respiratory rate, blood pressure, and age (Lim et al., 2003).

Table 2.2 maps each question in the survey according to their alignment within the micro-level constructs of CAF.

Table 2.2 Alignment of survey questions with CAF framework

Question category	Question in the survey	Alignments with CAF
Most useful CPRs	Select the top three clinical prediction rules that you believe are most useful in primary care. (Select up to three)	The question's alignment with the 'Usage Quality' (Use) aspect of the micro level reflects its aim to assess respondents' ability to identify the most applicable CPRs in their primary care settings, demonstrating their practical use and engagement.
Experience with CPRs	How much experience do you have using clinical prediction rules to either diagnose a disease or determine the prognosis of a disease during patient encounters in primary care?	This question aligns with the 'Usage Quality' (Use) aspect of the micro level in CAF, as it directly probes the extent of PCPs' engagement with CPRs in practice. The various response options (Significant, Moderate, Little, No experience) provide

Question category	Question in the survey	Alignments with CAF
		insights into their level of experience, indicating the extent of CPR usage.
The use of CPRs during patient encounters	To what extent do you agree that clinical prediction rules are easy for you to use during patient encounters in primary care?	This question is relevant to the 'Usage Quality' (Use) aspect of the micro level due to its assessment of user satisfaction and perceived ease of use, both of which are key components of usage quality. The range of responses (Strongly disagree, Disagree, Neither agree nor disagree, Agree, Strongly agree) measures the extent to which primary care providers find the rules user-friendly, directly impacting their willingness and ability to use them in clinical practice.
Impact of CPR Use in clinical practices	- To what extent do you agree that clinical prediction rules help you improve your diagnostic and prognostic accuracy during patient encounters in primary care? - To what extent do you agree that clinical prediction rules help you communicate disease diagnosis or prognosis during patient encounters in primary care?	These two questions relate to the 'Net Benefits' (Impact) aspect of the micro level. Both assess the perceived improvements in care quality and communication effectiveness, key elements of net benefits. The response options (Strongly disagree, Disagree, Neither agree nor disagree, Agree, Strongly agree) gauge the extent to which clinicians believe that CPRs contribute to better clinical outcomes and enhance their communication with patients, reflecting the tangible benefits of these tools in clinical practice.
Support to use CPRs	To what extent do you agree that your use of clinical prediction rules is well supported by your electronic health record (EHR)?	This question relates to the 'HIS Support Quality' (Support) at the micro level as it evaluates how effectively the EHR aids in applying CPRs. The response spectrum from Strongly disagree to Strongly agree measures the perceived adequacy of EHR

Question category	Question in the survey	Alignments with CAF
		support in clinical decision-making aided by CPRs.
Motivation	Select the three factors that motivate you most to use clinical prediction rules during patient encounters in primary care. (Select up to three)	Considering the options of this question, the question aligns with both the 'Usage Quality' (Use) and 'Net Benefits' (Impact) aspects of the micro level. The following is the breakdown of each option alignment.
	Improves diagnosis and prognosis accuracy	Net Benefits - Care quality (Impact)
	Improves the communication of diagnosis/ prognosis with patients	Net Benefits - Care quality (Impact)
	Saves time to diagnose/ prognose	Net Benefits – Productivity (Impact)
	Avoids unnecessary costs	Net Benefits - Cost-effectiveness (Impact)
	Easy to use	Usage Quality - User satisfaction (Use)
Barriers	Which of the following are barriers that you experience when using clinical prediction rules in primary care? (Select up to three)	Given the options presented in this question, it aligns with the HIS Support Quality', 'Usage Quality,' and 'Net Benefits' aspects at the micro level. The breakdown of each option's alignment follows:
	Disrupts clinical workflow	Net Benefits - Care quality (Impact)
	Do not improve diagnosis or prognosis accuracy	Net Benefits - Care quality (Impact)
	Concern of being sued	Net Benefits – Productivity (Impact)
	Gives results that are difficult to interpret	Net Benefits - Care quality (Impact)
	Time-consuming	Net Benefits – Productivity (Impact)
	Difficult to communicate the results to patients	Usage Quality – Use (Use)
	Interferes with clinician autonomy	Usage Quality - User satisfaction (Use)

Data Analysis

I analyzed the quantitative data using descriptive statistics to summarize the survey responses. The focus was on identifying the perception of use, support, and impacts of CPRs among PCPs in the WWAMI region. Using the R programming language (cran, n.d.), I employed descriptive statistics to summarize findings. In presenting and reporting the results, I displayed percentages from all responses comprehensively in tables for all questions.

Ethics approval

Participants received consent information through the recruitment emails. I clarified that they could access the survey only upon agreeing to participate. The Institutional Review Board (IRB) at the University of Washington in Seattle designated this study as exempt.

2.4 RESULTS

I conducted recruitment in September 2020, sending out two rounds of emails with a one-week interval between them. The target population within the WPRN consisted of 72 PCPs to whom I sent the recruitment emails. Out of these, I received 25 responses, yielding a response rate of 34.7%.

Participant characteristics

Table 2.3 shows the characteristics of respondents. Of the 25 respondents, the majority were M.D.s and had more than ten years of primary care experience.

Table 2.3 Participant characteristics

Characteristics	n (%)
Types of providers	N = 25
MD	20 (80%)
Non-MD (P.A., D.O., ARNP, and M.D. resident)	2 (8%)

Characteristics	n (%)
Declined to state	3 (12%)
Years of practices	N = 25
≤ 5 years	4 (16%)
> 5 - ≤ 10 years	7 (28%)
> 10 - ≤ 15 years	3 (12%)
> 15 - ≤ 20 years	2 (8%)
> 20 years	9 (36%)

Perception of use

Most useful CPRs

Table 2.4 presents the preferred CPRs as selected by respondents. CHADS2, for assessing stroke risk, and PHQ9/PREDICTNL, for depression, were the most chosen. The CENTOR Score, Framingham Score, and CURB Severity Score also ranked somewhat high in usefulness. The Bacterial Pneumonia Score and BCRAT models were less frequently chosen, while the remaining seven CPRs, including the Diabetic Foot Screen Tool and PredictAL, were not selected by any respondents.

Table 2.4 Most useful CPRs

CPR	n (%)
CHADS2	21 (84.0%)
PHQ9 Predict-NL	21 (84.0%)
Walsh Rule CENTOR Score	11 (44.0%)
Framingham Score	11 (44.0%)
CURB Severity Score	8 (32.0%)
Bacterial pneumonia score	1 (4.0%)

CPR	n (%)
Gail model/ BCRAT Nottingham Prognostic Index	1 (4.0%)
Diabetic Foot Screen Tool	0 (0%)
PredictAL	0 (0%)
Marburg Heart Score	0 (0%)
Flu score	0 (0%)
MICE Rule	0 (0%)
GERD Score	0 (0%)
Risk Assessment Tool (RAT) QCancer	0 (0%) 0 (0%)

Experience with CPRs

Table 2.5 displays the varied levels of experience among respondents in using CPRs. All participants reported having some experience, with the majority falling into the moderate experience category. A notable group indicated they had significant and moderate experience, while a smaller portion of respondents mentioned having only a little experience with CPRs.

Table 2.5 Level of experience with the use of CPRs

Type of experience	n (%)
Experiences with CPR	N = 25
Significant experience	4 (16%)
Moderate experience	14 (56%)
Little experience	7 (28%)
No experience	0 (0%)

The ease of use of CPRs during patient encounter

Table 2.6 shows the use of CPRs in clinical practice, focusing on their ease of use during patient encounters. About half of the respondents agree that CPRs are easy to use, while the other

half either disagree with this statement or remain neutral. Only a very small fraction strongly disagrees.

Table 2.6 The ease of use of CPRs during patient encounters

The use of CPRs	**n (%)**
Finding ease of use during patient encounters	N = 25
Strongly agree	0
Agree	12 (48%)
Neither agree nor disagree	4 (16%)
Disagree	8 (32%)
Strongly disagree	1 (4%)

Perception of support

Support for CPR use from EHR

Table 2.7 reveals that less than half of the respondents agree that they receive EHR support for using CPRs. This suggests a significant portion of respondents face challenges or do not perceive strong EHR support in their use of CPRs in clinical practice.

Table 2.7 Support for CPR use from EHR

Type of support on CPR use	**n (%)**
Having EHR Support for CPR Use	N = 25
Strongly agree	0
Agree	11 (44%)
Neither agree nor disagree	3 (12%)
Disagree	6 (24%)
Strongly disagree	5 (20%)

Perception of impact

Impact of CPR use in clinical practices

Table 2.8 describes the perceived impact of using CPRs, including improving accuracy and helping to communicate with patients. The majority of respondents agree that CPRs can improve diagnostic and prognostic accuracy. In terms of facilitating patient communication, over half agree, while about a third neither agree nor disagree. This finding suggests that CPRs are generally viewed as valuable tools for improving the accuracy of diagnoses and prognoses as well as supporting communication with patients.

Table 2.8 Perceived impact of CPR use

Types of impacts of CPR use	**n (%)**
Improving diagnostic and prognostic accuracy	N = 25
Strongly agree	0
Agree	19 (76%)
Neither agree nor disagree	1 (4%)
Disagree	1 (4%)
Strongly disagree	4 (16%)
Helping communication of diagnosis or prognosis with patients	N = 25
Strongly agree	3 (12%)
Agree	13 (52%)
Neither agree nor disagree	9 (36%)
Disagree	0
Strongly disagree	0

Motivation

Table 2.9 shows that the most frequently chosen motivating factor for using CPRs is their ease of use, aligning with 'Usage Quality' regarding user satisfaction in the CAF. Factors like

improving diagnostic/prognostic accuracy and avoiding unnecessary costs are also commonly selected, corresponding to the impact or 'Net Benefits' for care quality and cost-effectiveness.

Table 2.9 Motivation of the CPR use

Motivating factors to use CPRs	n (%)
Selected factors (select up to 3)	
Easy to use	16 (64%)
Improve diagnosis and prognosis accuracy	14 (56%)
Avoid unnecessary cost	12 (48.0%)
Improve the communication of diagnosis/ prognosis with patients	10 (40.0%)
Save time to diagnose/ prognose	8 (32.0%)
None of the above	0 (0%)
I don't have an opinion because I have no experience in using CPRs	0 (0%)

Barriers

Table 2.10 outlines the barriers experienced in using CPRs. The most reported barrier is the disruption of clinical workflow, aligning with the impact or 'Net Benefits', specifically Care Quality. Communication difficulties with patients, falling under impact or 'Net Benefits', is also notable barriers.

Table 2.10 Barriers experienced in using CPRs

Barriers experienced in using CPRs	n (%)
Selected factors (select up to 3)	
Disrupting Clinical Workflow	13 (52%)
Difficult to communicate with patients	12 (48.0%)
Time-consuming	8 (32.0%)
Gives results that are difficult to interpret	7 (28.0%)
Do not improve diagnosis and prognosis accuracy	4 (16.0%)
Interference with clinicians' autonomy	2 (8.0%)
Concerns of being sued	2 (8.0%)

Barriers experienced in using CPRs	n (%)
None of the above	0 (0%)
I don't have an opinion because I have no experience in using CPRs	0 (0%)

2.5 DISCUSSION

This study fills a research gap by assessing the perception of use, support and impact of CPR use among PCPs. Regarding the use, all respondents had experiences with CPRs and selected CHADS and PHQ as the most useful. Another key finding is that while ease of use was identified as the most motivating factor for CPRs, only about half of respondents agree that using CPRs during patient encounters is easy. For support, less than half of the respondents agreed that EHR effectively supports the use of CPRs. Additionally, respondents encountered barriers in communicating information derived from CPRs to patients, which is vital for ensuring the delivery of quality information. Regarding impact, the majority of respondents agreed that CPRs enhance diagnosis and prognosis accuracy and aid in communicating these with patients. Furthermore, these impacts, particularly accuracy improvement and avoiding unnecessary costs, seem to be among the main motivating factors for using CPRs. However, the main barriers identified include the disruption of clinical workflow and difficulties in communicating CPRs with patients.

Citation counts for the selected most useful CPRs

While observing the most useful CPRs favored by respondents in this study, I found differences compared to those identified in similar research targeting primary care clinicians. A study by Plüddemann et al. (2014) found that the Framingham Risk Score and PHQ9 were among the most preferred models, aligning closely with my findings. However, my findings diverge in the preference for CHADS2 and CENTOR Scores, which were highly favored by our respondents but not by Plüddemann et al. These discrepancies could be due to geographical differences—my

study being U.S.-based and Plüddemann's U.K.-based—and evolving clinical practices. These variations underscore the need for future research to identify the CPRs that clinicians find more useful in their practice.

Furthermore, the citation counts for the original pre-impact analysis studies for each CPR, as shown in Table 2.11, reveal that the citation count for these studies seems to correspond with their selection by respondents as the most useful CPRs. CPRs whose original studies are cited more frequently, such as CHADS2 and PHQ9/PREDICTNL, with thousands of citations, tend to be preferred as the most useful. In contrast, the Flu score, with only 41 citations, was not chosen as useful by any respondent. This observation suggests that CPRs with original studies that have a higher number of follow-up studies, as evidenced by a high citation count, may play an important role in their selection as the most useful in clinical practices by PCPs. The higher ranking of those CPRs might be attributed to those follow-up studies promoting CPRs to broader communities, including biomedical researchers and clinicians, and being incorporated into guidelines, thereby increasing clinician awareness and use of them in clinical practice.

Table 2.11 Citation counts of original studies for CPRs, ranked by CPRs selected as most useful in the survey

CPR	n (%)	Cited by (n)
CHADS2 (Gage et al., 2001)	21 (84.0%)	6132
PHQ9 (Spitzer et al., 1999) Predict-NL (Zuithoff et al., 2009)	21 (84.0%)	3,858,431
Walsh Rule (Walsh et al., 1975) CENTOR Score (Centor et al., 1981)	11 (44.0%)	1,951,081
Framingham Score (Wilson et al., 1998)	11 (44.0%)	12,365
CURB Severity Score (Lim et al., 2003)	8 (32.0%)	3,776

CPR	n (%)	Cited by (n)
Bacterial pneumonia score (Heckerling et al., 1990)	1 (4.0%)	286
Gail model/ BCRAT (Gail et al., 1989) Nottingham Prognostic Index (Haybittle et al., 1982)	1 (4.0%)	4,037, 897
Diabetic Foot Screen Tool (Murphy et al., 2012)	0 (0%)	39
PredictAL (King et al., 2011)	0 (0%)	12
Marburg Heart Score (Bösner et al., 2010)	0 (0%)	219
Flu score (Ebell et al., 2012)	0 (0%)	41
MICE Rule (Roalfe et al., 2012)	0 (0%)	40
GERD Score (Noya Horowitz et al., 2007)	0 (0%)	45
Risk Assessment Tool (RAT) (Hamilton, 2009) QCancer (Hippisley-Cox and Coupland, 2012)	0 (0%) 0 (0%)	272 138

Interestingly, this table may also indicate the reverse: clinicians pick the most useful CPRs based on the evidence directly from the research papers. Once clinicians identify these tools, biomedical researchers become aware of their practical utility and conduct follow-up studies. This reverse process shows that the initial clinical use of CPRs can drive academic interest and further research, reinforcing the clinical relevance of these tools through additional studies and citations.

However, clinicians often do not rely solely on research papers to use tools in their practice. Instead, they use a combination of guidelines, research papers, and other online resources (Davies, 2007; Keylen et al., 2020). If CPRs appear more frequently in these various resources, clinicians are more likely to become aware of them and try them in clinical practice. This indicates that CPRs

with more follow-up studies may have a greater chance of being included in multiple resources, thus enhancing their visibility and perceived utility among clinicians.

The use of CPR and its motivating factors and barriers

In this study, all participants had experience using CPRs in their clinical practices, as indicated by none mentioning that they had no experience with CPRs. To explore more about this use, a follow-up question was asked about the motivating factors that drive the use of CPRs. Among the top four factors, pragmatic considerations such as ease of use ranked number one, and avoiding unnecessary costs was the third most chosen factor.

The key factor regarding CPR use, 'ease of use,' aligns with other studies that emphasize 'ease of use' as a critical factor in determining the usefulness of CPRs (Richardson et al., 2015). However, it is intriguing that less than half of the respondents agree that using CPRs during patient encounters is easy. This discrepancy may be attributed to varying levels of difficulty in the application of CPRs, HIS quality support, as observed in the EHR supports, or direct influences from key aspects in the meso levels of the CAF, which were not part of this study. Therefore, these observations warrant further research to better understand and address this gap such as knowing if their organizations are promoting use or if it is an individual choice. This understanding may help in finding ways that work for both individual PCPs and their organizations, aiming to optimize the adoption and effective use of CPRs in primary care settings.

The main factors that may contribute to the motivating factors can also be seen in the responses about barriers. The top-ranked barrier was the disruption of clinical workflows while using CPRs, followed by difficulties in communicating with patients. These two barriers are also common in other studies (Kennedy and Gallego, 2019). Addressing these barriers warrants further research to effectively implement CPRs in clinical practice. To ensure that CPRs are easy to use,

they should not disrupt clinical workflows and should facilitate, rather than hinder, communication with patients (Kappen et al., 2016).

The support for CPR use

With less than half of the respondents agreeing that EHR effectively supports the use of CPRs, this observation indicates challenges in the support of CPR use. Integrating clinical decision support like CPRs into EHR systems can be highly beneficial, as it can enhance the quality of information and communication with patients (Sutton et al., 2020). This integration not only improves accessibility to the application of CPRs but also streamlines the communication process, ensuring that both clinicians and patients have a better understanding of their diagnosis and prognosis as well as follow-up disease management.

The impacts on CPR use

Regarding impact, the majority of respondents perceived potential benefits of CPRs, citing the improvement of diagnosis and prognosis accuracy as a second primary motivator. This aligns with existing literature that claims CPRs improve diagnosis and prognosis accuracy (Barnett et al., 2019; Kostopoulou et al., 2015). Therefore, the motivation to use CPRs in clinical settings is supported by evidence demonstrating their effectiveness in improving diagnosis and prognosis accuracy.

One key impact acknowledged by respondents yet identified as a notable barrier is aiding in communicating diagnoses and prognoses with patients. This impact may come through facilitating the translation of complex medical data into more understandable terms, allowing patients to better comprehend their health situation (Walsh et al., 2021). Such clarity can promote shared decision-making, enabling patients to be more informed and actively participate in treatment discussions. While CPRs can simplify medical information and build trust in the patient-

provider relationship, this study also found barriers to conveying CPR-derived information to patients. To address these barriers, activities such as further education embedded in the medical education curriculum or continuing professional education may be considered.

The other impact identified as a prominent barrier to using CPRs by respondents is the disruption to clinical workflow. This disruption can occur in several ways, such as challenges integrating CPRs with current systems like EHR. Additionally, using CPRs may require extra time for data entry, analysis, or interpretation, which can be particularly taxing during busy clinical schedules. Moreover, the need for training and familiarity with these tools can further slow down workflow, as healthcare providers might need additional time to effectively understand and apply CPR in their practice.

Limitations

Despite the contributions of this study, some limitations warrant consideration. The convenience sample of 25 participants, while potentially reflective of the target population within the WPRN, may limit the generalizability of findings. Recruitment challenges and the survey's focused scope, constrained to 10 questions, also restricted the depth of exploration into the complex factors affecting CPR use among primary care providers and incomplete mapping into CAF constructs. A larger sample and a more comprehensive survey could provide a deeper understanding of participant characteristics, such as detailed demographics and practice experiences that may influence CPR use. Additionally, a more extensive survey would enable a deeper exploration of relationships between variables and facilitate subgroup analyses based on level of experience using CPRs or respondent characteristics (i.e., years of practice, or types of providers). These limitations should be kept in mind when interpreting the findings.

Future studies

Future studies based on this research could explore several areas:

1. One approach is to use the CAF with more robust instruments. This method would allow for a thorough evaluation of CPRs across the micro, meso, and macro levels of CAF. Employing detailed surveys and observational studies could capture diverse data ranging from user satisfaction to systemic integration challenges, including the broader impacts on organizational structures and policies. This approach aims to develop a comprehensive understanding of the complexities of implementing CPRs in clinical settings.

2. Another valuable direction for future research is expanding the scope of respondents. Reaching out to a wider range of PCPs from diverse networks could enhance the study's representativeness and provide more generalized insights. This could involve targeting different geographical regions or practice settings to understand the varied experiences and perceptions of PCPs regarding the use of CPRs.

3. The third area of focus could be on longitudinal studies to observe changes over time in the adoption and impact of CPRs. This would help in understanding how training, policy changes, or technological advancements influence the long-term effectiveness and integration of CPRs in clinical practice. Such studies could also explore the evolving challenges and opportunities in implementing CPRs, providing valuable data for continuous improvement in primary healthcare.

4. Research on CPMs that inform CPRs should focus on both accuracy and utility. These studies need to balance the research objectives often centered on accuracy with the practical challenges clinicians encounter regarding usefulness or utility. This approach would provide a more holistic understanding of CPMs in clinical practice, addressing the needs and constraints of healthcare providers in real-world scenarios.

5. Studies that focus on enhancing the quality of preimpact analysis studies that develop and validate CPMs and inform CPRs. My study findings indicate that CPRs with original studies that have garnered a higher number of follow-up studies, as evidenced by a high citation count, influence their selection as the most useful in clinical practice by PCPs. However, the current state of the reporting quality of preimpact analysis studies remains poor, according to biomedical researchers (Damen et al., 2016; Dhiman et al., 2021; Heus et al., 2018; Najafabadi et al., 2020; Yang et al., 2022). Therefore, an urgent step is to promote the improvement of preimpact analysis studies to ensure that biomedical researchers can be expected to conduct follow-up studies that support clinicians in recognizing the utility of these CPMs. This necessitates engaging biomedical researchers to understand the challenges and needs related to ensuring the quality of these preimpact analysis CPM studies, thereby facilitating the development of more useful CPRs for clinical practice. I selected this study direction for my dissertation.

2.6 CONCLUSION

In conclusion, this study offers insights into the use, support, and impact of CPRs in primary care settings. Notably, all participating clinicians reported some level of experience with CPRs and were able to rate some CPRs as the most useful. A majority of respondents agreed that using CPRs can improve diagnostic and prognostic accuracy, reaffirming that studies on prediction models for clinical purposes remain critical. Furthermore, improving diagnosis and prognosis accuracy ranks second in motivating the adoption of CPRs, following the top motivator of ease of use, as a reason to adopt CPRs in clinical practice. These findings confirm that PCPs continue to find these tools useful and are incorporating CPRs into their practice as they continue to experience CPR effectiveness in primary care practices.

Moreover, the citation counts for these studies appear to align with the respondents' selection of the most useful CPRs. Frequently cited CPRs tend to be preferred as the most useful, indicating that those with extensive follow-up studies are more likely to become known and gain traction for use in clinical practices. This dynamic suggests that efforts to enhance the reporting quality of CPM study results in research papers may provide a foundation that supports follow-up studies, thereby facilitating the ongoing use and refinement of CPRs in clinical practice. Continued improvements in the quality of preimpact analysis studies are essential, ensuring they are adequately disseminated and meet the practical needs of clinicians who rely on these tools daily.

Chapter 3. USING PUBMED TO BOLSTER THE RECRUITMENT OF BIOMEDICAL RESEARCHERS

3.1 INTRODUCTION

Biomedical researchers have raised concerns about the quality of CPMs in research papers, as highlighted in recent systematic review studies (Damen et al., 2016; Dhiman et al., 2021; Heus et al., 2018; Yusuf et al., 2020). CPMs are beneficial in clinical practice, aiding clinicians in improving diagnostic and prognostic accuracy (Barnett et al., 2019; Kostopoulou et al., 2015). The issues identified include incomplete or selective reporting of significant study methods and results, misinterpretation of results, and misleading data visualization representations (ESHRE Capri Workshop Group, 2018; Simera et al., 2010; Weissgerber et al., 2015). These issues lead to poor reporting, adversely affecting the quality attributes expected in CPM study reports that develop and validate CPMs, including transparency, understandability, usefulness, and trustworthiness (Damen et al., 2016; Dhiman et al., 2021; Heus et al., 2018; Yusuf et al., 2020).Biomedical researchers have raised concerns about the quality of CPMs in research papers, as highlighted in recent systematic review studies (Damen et al., 2016; Dhiman et al., 2021; Heus et al., 2018; Yusuf et al., 2020). CPMs are beneficial in clinical practice, aiding clinicians in improving diagnostic and prognostic accuracy (Barnett et al., 2019; Kostopoulou et al., 2015). The issues identified include incomplete or selective reporting of significant study methods and results, misinterpretation of results, and misleading data visualization representations (ESHRE Capri Workshop Group, 2018; Simera et al., 2010; Weissgerber et al., 2015). These issues lead to poor reporting, adversely affecting the quality attributes expected in CPM study reports that develop

and validate CPMs, including transparency, understandability, usefulness, and trustworthiness (Damen et al., 2016; Dhiman et al., 2021; Heus et al., 2018; Yusuf et al., 2020).

Reporting guidelines aim to enhance the quality of research papers, including those on CPMs. Notable guidelines are TRIPOD (Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis) (Collins et al., 2015) and PROBAST (Prediction model Risk Of Bias ASsessment Tool) (Wolff et al., 2019). These guidelines, especially TRIPOD and PROBAST, primarily focus on issues like incomplete and selective reporting and emphasize transparency. However, they often overlook crucial issues like the misinterpretation of results and misleading data visualizations, which affect whether CPM study results are understandable. Furthermore, ensuring that results are useful and trustworthy also remains neglected in these reporting guidelines.

The development of guidelines for quality reporting in research papers, including TRIPOD and PROBAST, typically follows general steps for creating health research reporting guidelines (Moher et al., 2010). These steps usually involve conducting literature reviews and employing the Delphi method, a structured consensus-building process among domain experts. This approach often focuses on a specific group of experts, usually recruited through professional networks, which might exclude a broader range of biomedical researchers. Recruiting through mailing lists is a common means to engage participants in such studies that involve domain experts (Norris et al., 2021). This limitation raises concerns about the guidelines' comprehensiveness in addressing the diverse challenges among domain experts of achieving the desired qualities in CPM studies.

PubMed provides information that allows researchers to engage biomedical researchers as domain experts in studies aimed at improving the quality of research papers (Khalifa, 2019). PubMed contains records of authorships, including email addresses of corresponding authors of

articles, based on information provided by academic journals listed in the PubMed databases ("Authorship in MEDLINE," 2023). However, evidence of recruitment effectiveness in engaging a broad range of biomedical researchers as domain experts using PubMed records to improve the quality of research papers remains absent.

3.2 STUDY OBJECTIVES

This study evaluates an alternative approach to recruiting biomedical researchers using PubMed records compared with recruitment through traditional mailing lists sent to professional networks. PubMed, a widely used database of medical literature, contains records with contact information (i.e., emails) of biomedical researchers who have authored research papers, including CPMs (Khalifa, 2019). I leveraged this information as a recruitment source. The aim is to assess whether this approach can effectively reach a larger and more diverse sample of biomedical researchers than mailing lists of professional networks, which could improve the diversity of perspectives on improving the quality of CPM research papers.

3.3 METHODS

Eligibility criteria

Eligibility criteria included having experience as an author or reviewer of CPM research papers in peer-reviewed biomedical journals or conference proceedings and being 18 years or older.

Recruitment strategy

To compare between recruiting participants for a study about improving the quality of CPM research papers using email addresses obtained from PubMed records and with mailing lists, I followed five steps: 1) asking colleagues about mailing list of biomedical researchers with

experiences in authoring or reviewing research papers about CPMs, 2) sending recruitments through professional mailing lists, 2) obtaining PubMed records of articles about CPMs and their corresponding contact information using a systematic search, 3) extracting email addresses contained in the PubMed records, 4) sending recruitments to email addresses collected from PubMed records, and 5) collecting and analyzing data from the recruitment strategy. The recruitment email for both using mailing lists and PubMed records contained a REDCap link that directed email recipients to participate in my survey. The details of each recruitment strategy are as follows.

Mailing list recruitment: The mailing list groups that I targeted for my survey recruitment should have members who are professional biomedical researchers with experience in CPM studies, either as authors or reviewers of CPM research papers. Through information provided by my networks (i.e., peers, faculty members at the University of Washington, and Ph.D. committees), I identified three professional mailing lists: 1) a mailing list with members of biomedical researchers from primary care researchers in the United States and Europe, 2) biomedical researchers from a hospital in Seattle, and 3) academic faculty from a University in Seattle.

For this mailing list recruitment, I sent the recruitment email to the list's administrator. After approval, the administrator sent the recruitment email to the mailing list. Consequently, I could not track information such as the number of biomedical researchers who received the recruitment email. The only data I could collect was the number of responses to recruitment, as I used separate REDCap survey links for each recruitment method.

PubMed email recruitment: The collection of email addresses through PubMed records began by setting up a query strategy, implementing the query on the PubMed database, and downloading the query results using PubMed format. The search strategy was limited to studies

related to CPMs aimed for use in primary care between 2015 and October 2022. The full keyword strategies are in Appendix B. From the PubMed records, I obtained the email addresses of the authors of the CPM research papers.

For the PubMed recruitment method, I used Gmail to send and track the status of email delivery. To send the recruitments to emails from the PubMed record, I used Gmail app scripts to send emails in bulk ("Create a mail merge with Gmail & Google Sheets | Apps Script," n.d.). I sent emails to each address twice. In the second email, I acknowledged that I would appreciate it if they responded. For those who have not yet responded, I kindly ask that they consider filling it out.

PubMed data collection

The first data collection was obtained by extracting PubMed records, which included standard data elements such as PMID, title, journal title, publication date, authors, countries of authors, and email addresses ("MEDLINE/PubMed Data Element (Field) Descriptions," n.d.). The second dataset came from tracking the emails I sent to determine if they were received by respondents. I labeled each email sent and its status—such as 'sent,' 'auto-response,' or 'undelivered'—with 'undelivered' indicating emails that did not yield any response after being sent twice. A customized Gmail app script was used to collect this data.

Data analysis

The first data analysis involved calculating numbers while doing the recruitment, including the number of articles, authors, emails sent, and emails that reached the targeted respondents, as well as the response rate. The second analysis explored patterns related to journal titles with the most publications and email addresses to identify any discrepancies. This helped determine whether email addresses were obtained from all journals. The third analysis examined the

geographic distribution of email addresses in relation to the authors' countries of affiliated institutions. All these analyses were conducted using R.

Ethics approval

In the IRB application for my study recruitment submitted to the University of Washington's Institutional Review Board (IRB) in Seattle, I specified that my recruitment strategy would involve using publicly published email addresses of authors from academic journal articles and mailing lists related to CPM studies. The IRB granted my study an exempt status based on this approach.

3.4 RESULTS

The recruitment for this study spanned from mid-November 2022 to mid-January 2023. I employed two methods for recruitment: mailing lists and PubMed records, with a three-week interval between the two approaches.

Recruitment and response rate

Figure 3-1 provides a summary of the recruitment strategies for both mailing list and PubMed recruitment, along with numbers from each step and the final response rate. For the mailing list recruitment, I received 12 responses. It is worth noting that calculating a response rate for this method was not feasible, as the total number of list members was unknown.

On the other hand, the PubMed email recruitment was more quantifiable. I initially identified 9,550 PubMed records, which corresponded to 43,534 unique authors and 6,860 unique email addresses for corresponding authors. Recruitment emails were sent to these 6,860 addresses. Of the 6860 sent, 1,921 emails (28%) were not delivered for various reasons. These included 530 emails with an 'undelivered' status, 13 with an 'auto-response' status, and 1,378 that were 'missing,'

meaning they had no status reports from Gmail. Furthermore, out of the 6860 sent, 4,939 emails were successfully delivered (72%), resulting in 251 survey responses—a response rate of 5%.

.

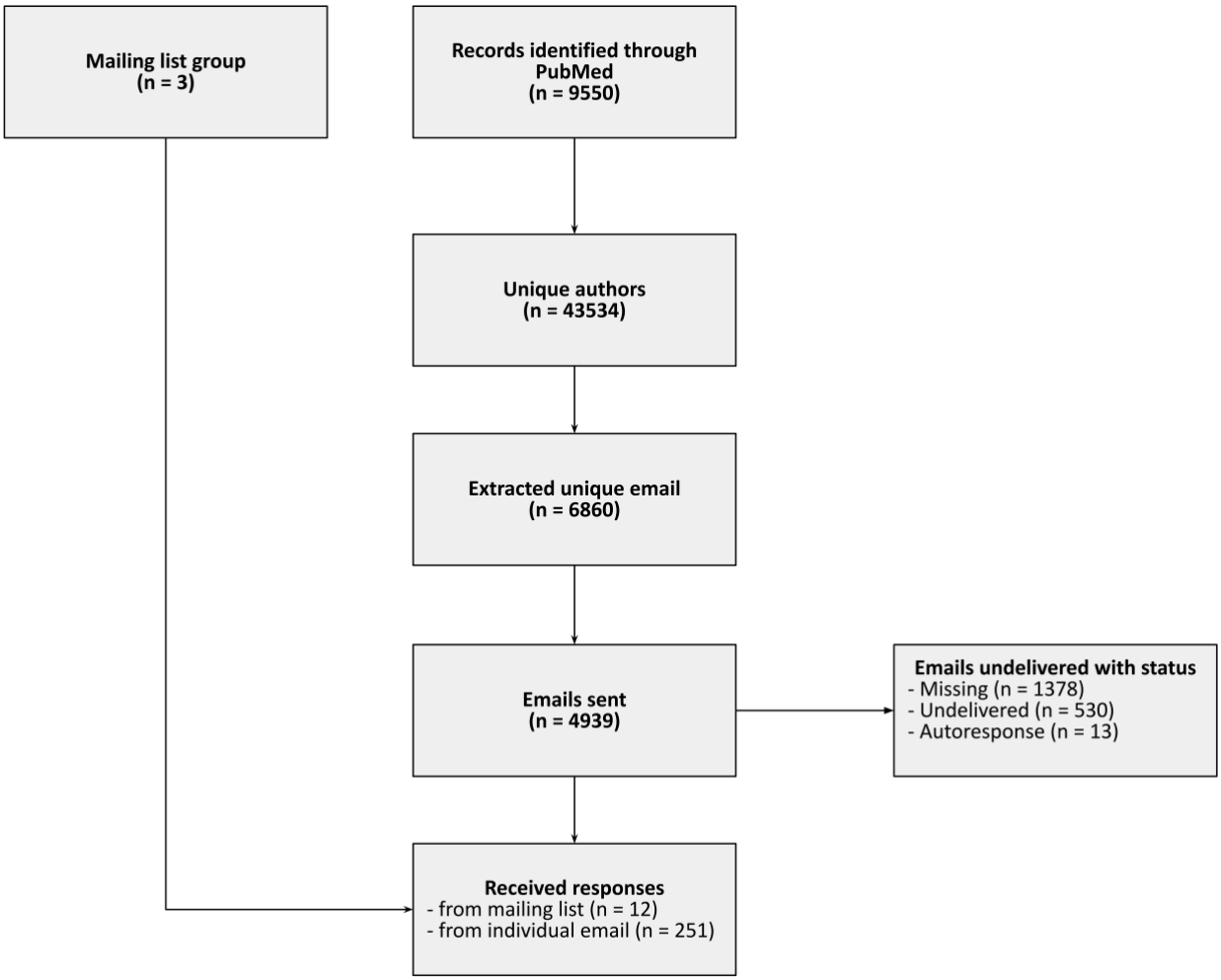

Figure 3-1 Recruitment strategy

Journals with email addresses

Using PubMed records allowed me to identify the number of unique journal titles that I obtained and which journals have email addresses. My analysis of journal titles identified 1,900 unique journal titles, yet only 966 (51%) included author email addresses. Table 3.1 demonstrates this discrepancy by showcasing the top 10 journals with the most email addresses. Notably, the

rankings differ between the two categories, the top 10 journals, with the most publications and email addresses indicating the possibility that not all journals or articles in a journal include email addresses for authors. For example, while 'PLOS ONE' had the most publications with 377, it did not include email addresses. 'Scientific Reports' provided the most email addresses with 313.

Table 3.1 Journal titles with most publications, some of which do not provide email addresses for authors

Journal title with the most publication	**n**	**Journal title with the most email addresses**	**n**
PLOS ONE	377	Scientific reports	313
Scientific reports	265	International Journal of Infectious Diseases	188
BMJ open	179	Journal of Stroke and Cerebrovascular Diseases	133
Journal of the American Heart Association	103	BMJ Open	120
International Journal of Environmental Research and Public Health	84	International Journal of Cardiology	103
International Journal of Cardiology	77	BMC Infectious Diseases	86
Medicine	69	Thrombosis Research	86
Stroke	68	Osteoarthritis and Cartilage	80
Journal of Stroke and Cerebrovascular diseases	64	Resuscitation	78
BMC Medical Informatics and Decision Making	59	The American Journal of Emergency Medicine	71

Geographic distribution

In the geographic analysis, I identified email addresses for authors' countries of affiliated institutions from as many as 124 different countries across five country regions (i.e., Europe, Asia, America, Oceania, and Africa). I used United Nations references to categorize countries into five regions (United Nations, n.d.).

Table 3.2 provides a breakdown of the distribution of regions with the most email addresses according to the authors' country of affiliated institutions. Notably, Europe emerges as the leading

region, representing 42.9% of email addresses with a total count of 4503. Following Europe, Asia contributes 25.8% with 2709 email addresses, while America accounts for 23.8%, totaling 2501 email addresses. Oceania and Africa are represented by 466 (4.4%) and 327 (3.1%) email addresses, respectively. Please note that the total counts of email addresses in this table differ from those presented in Figure 3-1. The figures include only unique email addresses. In contrast, this table counts email addresses associated with one or more countries.

Table 3.2 Distribution of regions with the most email addresses and participant origin countries

Country	**Email addresses n (%)**
	N = 10506
Europe	4503 (42.9%)
Asia	2709 (25.8%)
America	2501 (23.8%)
Oceania	466 (4.4%)
Africa	327 (3.1%)

3.5 DISCUSSION

This study demonstrates that leveraging PubMed records can serve as an alternative recruitment tool to bolster the reach over traditional mailing lists. This approach also enabled me to apply further analysis, including calculating response rate and providing evidence that recruiting a large number of diverse biomedical researchers from different regions for a survey to improve the quality of research papers is feasible.

The study revealed that a notable number of journals do not include author email addresses in their PubMed records, creating a potential obstacle for using these records in researcher recruitment but preserving the privacy of authors. Upon examining the policies of some journals, it was evident that there are no specific mandates for the inclusion of author email addresses. My

further examination using R, from the top 10 list of journals with the most publications, journals without email addresses listed are 1) 'PLOS ONE,' and 2) 'Medicine.' The remaining eight journals have listed the email addresses of the authors. Both groups of journals, whether they include email addresses or not, are recognized as reputable and high-ranking in their respective fields ("Scimago Journal & Country Rank," n.d.).

This lack of consistency in not including email addresses in PubMed records from a number of respectable journals could affect the representativeness of biomedical researchers in studies that apply this recruitment method. The absence of email addresses from specific journals, especially those that focus on particular subjects or are prevalent in certain regions, could lead to inadequate representation. Consequently, researchers with interests typically published in these journals or regions may be underrepresented in studies that depend on PubMed for recruitment.

The results of this study indicate that using PubMed email records is a promising approach for recruiting biomedical researchers for studies aimed at improving the quality of CPM research papers and potentially other topics. The 5% response rate may be low compared to other studies that recruit professionals from healthcare backgrounds (Meyer et al., 2022). However, the response rate of this survey falls within the range of 5-10%, the minimum requirement specified by the National Survey of Student Engagement (NSSE) for survey data to be considered reliable (Fosnacht et al., 2017). Furthermore, the large number of unique email addresses obtained through the PubMed search demonstrates the potential scale of this approach.

Limitation

While the method of using PubMed email records successfully reached researchers from regions with high publication rates, several limitations must be acknowledged. First, the absence of email records for authors in approximately 50% of journals limits the comprehensiveness of the

recruitment strategy. This limitation is compounded by a high rate of undelivered emails due to unintended errors when using Gmail services to send recruitment emails in bulk, which further constrains the method's efficacy. Future research should further explore journal policies regarding the inclusion or exclusion of email addresses in PubMed records and understand the policies of the Gmail service surrounding sending emails in bulk to better navigate these limitations.

An additional ethical consideration I must address is the potential vulnerability of PubMed to email address harvesting, as highlighted by Brendan Thomas (Thomas, 2011). While PubMed serves as a valuable resource for academic research, its structure allows for the easy retrieval of authors' email addresses, raising concerns about privacy and the potential for spam. I acknowledge the controversy surrounding the use of these email addresses for recruitment purposes. However, it is worth noting that the intent aligns with academic research goals rather than commercial or malicious use. The use of email addresses from PubMed records was aimed at fostering inclusivity by reaching a diverse and representative sample of biomedical researchers, particularly those who may not be accessible through traditional professional networks or mailing lists.

Future studies

Future studies can benefit from adopting the recruitment methodology demonstrated in this study, extending its application beyond the context of CPM studies to encompass a broad range of topics relevant to biomedical researchers. By utilizing PubMed email records as a recruitment tool, researchers can reach a more diverse and representative sample of the biomedical research community.

My experience also has shown that the sender's credentials can significantly impact the response rate. As a Ph.D. student sending recruitment emails from a Gmail account, I received emails that asked for my credentials. Future studies could involve using research organization

email accounts and more established figures in the field as the primary point of contact or include endorsements from reputable institutions or researchers in the recruitment materials. This approach could foster greater trust among potential participants, improving response rates and more robust engagement.

3.6 CONCLUSION

In conclusion, this study demonstrates that leveraging PubMed email records is a viable recruitment strategy for engaging a diverse sample of biomedical researchers compared to using traditional mailing lists. Such recruitment can potentially improve the quality of CPM research papers by engaging a broader representation of biomedical researchers. While this approach yielded a response rate within the acceptable range defined by the National Survey of Student Engagement (NSSE) (Fosnacht et al., 2017), it also revealed significant limitations. These limitations include the absence of email records in approximately half of the journal titles and a high rate of undelivered emails due to unintended errors, especially those emails sent in bulk without status reports confirming delivery. To address this issue, it is important to understand better Google's email service policies for bulk email delivery or explore alternative email services with clearer policies and better record tracking Despite these challenges, the recruitment method shows promise in reaching a broad and geographically diverse audience, fulfilling the study's aim to assess the strengths and weaknesses of this recruitment approach. Future research should address these limitations to further optimize the utility of PubMed email records as a recruitment tool in biomedical research.

Chapter 4. ADMINISTERING A MIXED-METHOD SURVEY TO IDENTIFY CHALLENGES EXPERIENCED BY BIOMEDICAL RESEARCHERS IN PRESENTING UNDERSTANDABLE, USEFUL, AND TRUSTWORTHY RESULTS OF CLINICAL PREDICTION MODEL STUDIES

4.1 INTRODUCTION

Despite the growing number of publications, the biomedical research community found that poor reporting quality of study results for CPM research papers that report preimpact analysis studies remain common (Damen et al., 2016; Dhiman et al., 2021; Heus et al., 2018; Najafabadi et al., 2020). The community identifies that the quality characteristics of results for the preimpact analysis studies in peer-reviewed research papers should be transparent, understandable, useful, and trustworthy (Damen et al., 2016; Dhiman et al., 2021; Heus et al., 2018; Yusuf et al., 2020). The poor reporting quality of the CPM study results could lead to difficulties among biomedical researchers in following the studies, such as conducting further studies for replication, validation, or systematic review and adopting the models in clinical practices (Dhiman et al., 2021; Maiga et al., 2019).

The biomedical research community developed guidelines to help researchers produce high-quality CPM research papers for preimpact analysis studies, such as TRIPOD (Moons et al., 2015), PROBAST (Wolff et al., 2019), and Guidelines for Developing and Reporting Machine Learning Predictive Models in Biomedical Research (Luo et al., 2016). However, these guidelines

primarily emphasize the production of research papers that are complete and transparent, overlooking the other three quality attributes of understandable, useful, and trustworthy results of CPM studies (Damen et al., 2016; Dhiman et al., 2021; Heus et al., 2018; Yusuf et al., 2020).

Furthermore, the development process for these guidelines has historically relied on recruiting peer researchers through professional networks, with no preceding studies assessing challenges faced by biomedical researchers in producing high-quality CPM papers, particularly in achieving the transparency attribute as intended by the TRIPOD guideline (Collins et al., 2015; Wolff et al., 2019). The development of the guideline through recruiting peer researchers from professional networks risks exclusivity, potentially limiting the perspectives of a broader range of researchers. As a consequence of the absence of preceding studies that assess these challenges, the extent to which these guidelines address actual challenges experienced by the wider biomedical research community to improve the quality of CPM research papers remains uncertain.

4.2 STUDY OBJECTIVE

The current practice of improving the quality of research papers, such as through developing reporting guidelines often excludes identifying the challenges faced by biomedical researchers. This study emphasizes that attempts to improve the quality of CPM study results in research papers across three quality attributes—understandable, useful, and trustworthy—should begin with understanding these challenges. These challenges could encompass multiple perspectives, as biomedical researchers can assume various roles, such as authors, or reviewers of CPM research papers. Thus, the overall objective of this study is to address the gap by identifying the challenges experienced by biomedical researchers in perceiving and producing results of CPM studies that are understandable, useful, and trustworthy as authors and reviewers. The study has four specific objectives:

1) **To determine the extent to which biomedical researchers perceive the study results of CPMs as understandable, useful, and trustworthy.** Research questions (RQ) include:

   RQ1a: What is the frequency in which biomedical researchers perceive the results of CPM studies as understandable, useful, and trustworthy?

   RQ1b: Do the perceptions of biomedical researchers of CPM study results differ across the three quality attributes: understandable, useful, and trustworthy?

   RQ1c: Do perceptions of CPM study results as understandable, useful, and trustworthy vary by biomedical researchers' characteristics?

2) **To quantitatively characterize the challenges experienced by authors of CPM papers.** This objective aims to answer the following research questions:

   RQ2a: What is the level of difficulties that biomedical researchers as authors experience when producing understandable, useful, and trustworthy results in CPM studies?

   RQ2b: Do the level of difficulties experienced by biomedical researchers as authors in producing understandable, useful, and trustworthy results of CPM studies differ across the three quality attributes?

   RQ2c: Do experiences encountering difficulties in authoring understandable, useful, and trustworthy results vary by biomedical researchers' characteristics?

3) **To quantitatively characterize the challenges experienced by reviewers of CPM papers.** This objective aims to answer the following research questions:

   RQ3a: What is the level of difficulties that biomedical researchers as reviewers experience when providing peer review feedback to ensure understandable, useful, and trustworthy results in CPM studies?

RQ3b: Do the level of difficulties experienced by biomedical researchers as reviewers in providing feedback to ensure understandable, useful, and trustworthy results of CPM studies differ across the three quality attributes?

RQ3c: Do experiences in encountering difficulties in reviewing understandable, useful, and trustworthy results vary by biomedical researchers' characteristics?

4) **Objective 4 is to qualitatively describe the reasons biomedical researchers perceive the challenges they face when authoring and reviewing results that are understandable, useful, and trustworthy.** This objective aims to answer the following research questions:

RQ4: What are the reasons underlying challenges that respondents experience?

4.3 METHODS

Study design

Using a mixed-method online survey, the purpose of this study is to understand the challenges biomedical researchers face when presenting results in CPM studies. Both quantitative and qualitative data collection informed the data analysis. The quantitative analysis measures respondents' perceptions of the results in CPM studies as understandable, useful, and trustworthy, including difficulties experienced from the perspectives of biomedical researchers. The qualitative analysis investigated the reasons behind these difficulties.

Recruitment

This study recruited biomedical researchers as respondents by sending email addresses through professional mailing lists and using email addresses obtained from PubMed records. The details of the recruitment are described in Chapter 3. Additionally, participants received 25 USD

for completing the interview, which was provided by The Ira Kalet and Fred Wolf Endowment Fund.

Data collection

I implemented the survey and collected the data using REDCap, a secure web application designed to support data capture for research studies ("REDCap," n.d.). This study centered around three quality attributes of results in CPM studies: understandable, useful, and trustworthy. **Understandable study results** refer to the degree to which a reader is able to interpret, exemplify, classify, summarize, infer, compare, and explain the study results of prediction model research papers (Anderson and Krathwohl, 2001); **Useful study results** refer to the degree to which a biomedical researcher is satisfied with their perceived achievement of pragmatic goals, including the goals after reading the study results of prediction model research papers (IEC, 2011); **Trustworthy study results** refer to the degree to which a reader has confidence that the study results of prediction model research papers provide information as they should (IEC, 2011).

The quantitative section of the survey consisted of questions organized into five categories: 1) demographic information, including gender (man, woman, fill in), age (18-24, 25-34, 35-44, 45-65, 66-79, 80+ years old), organizational affiliation (i.e., academia, industry, government, consultancy, freelance, other) , and country of origin (fill in); 2) professional experience, including the number of CPM publications (i.e., <1, 1–5, 6–10, 11–20, > 20 years) and years of experience as a author and/or a reviewer (i.e., i.e., <1, 1–5, 6–10, 11–20, > 20 publications); 3) the frequency with which respondents perceive study results of CPM research papers to be understandable, useful, and trustworthy (5 point Likert scale: 1 'always', 2 'often', 3 'sometimes', 4 'rarely', 5 'never'); and 4) the level of difficulty respondents as author and reviewer encounter in ensuring that CPM

studies yield understandable, useful, and trustworthy results (5-point Likert scale: 1 'very easy', 2 'easy', 3 'neutral', 4 'difficult', 5 'very difficult').

The qualitative portion of the survey employed open-ended questions to explore respondents' reasoning about the difficulties reported as authors and reviewers (category 4). This qualitative portion aimed to understand the challenges they face as authors or reviewers in ensuring understandable, useful, and trustworthy results of CPM studies. See Appendix C for the complete survey instrument.

Prior to distributing the survey, I conducted a pilot with five colleagues to validate whether the survey instrument would be able to capture the intended study objectives. In this pilot, they provided feedback, focusing on improving the clarity, relevance, and comprehensiveness of the survey.

Quantitative data analysis

I summarized the demographics and respondents' experiences in authoring and reviewing CPM studies with descriptive statistics. For country of origin, I categorized them into regions based on United Nations references to region categorization (United Nations, n.d.). I employed a multi-method approach to address the study's specific research questions (RQs). The statistical techniques used included a) descriptive statistics to summarize respondent characteristics, b) Friedman tests to compare the differences in perceptions among biomedical researchers regarding the frequency of perceiving CPM study results as understandable, useful, and trustworthy, as well as the level of difficulty in ensuring that CPM studies yield understandable, useful, and trustworthy results, (Marshall and Marquier, n.d.), and c) multivariate logistic regression to examine the associations between biomedical researchers' perceptions of the understandable, useful, and trustworthy results of CPM studies and their demographic characteristics, as well as their

experiences as authors and reviewers. (Sperandei, 2014). All analyses were conducted using the R programming language (cran, n.d.).

Descriptive statistics: For RQ1a, RQ2a, and RQ3a, I analyzed the frequency with which biomedical researchers find the results of CPM studies understandable, useful, and trustworthy, as well as the frequency with which they experience difficulties when authoring or providing peer review feedback to ensure such results in CPM studies. All frequencies were reported in their original Likert scale values.

Friedman test and Nemenyi post hoc test. The Friedman test is a non-parametric statistical method used to detect differences in ranks across multiple related samples (Mangiafico, 2016). The analysis aimed to ascertain variances in 1) how often respondents found (RQ1b), 2) the difficulty among authors (RQ2b) in producing, and 3) the challenge for reviewers in providing feedback to ensure (RQ3b) that the results of CPMs are understandable, useful, and trustworthy. The outcome of the test indicates whether there is a statistically significant difference across the three evaluated quality attributes (i.e., understandable, useful, and trustworthy). Nemenyi post-hoc tests were employed to investigate differences in paired comparisons. This method is particularly appropriate for identifying specific pairs of groups with significant rank differences after the Friedman test, which indicated an overall significant difference between the pair of the three quality attributes.

Multivariable logistic regression: Lastly, multivariable logistic regression was used to address RQ1c, RQ2c, and RQ3c. I binarized the 5-point Likert scale responses for the outcome variables into two groups. For RQ1c, the newly categorized groups were 'frequent' (i.e., always and often responses)' and 'less frequent' (i.e., sometimes, rarely, never responses). The rationale behind this binarization was to better align with the study's objectives, which assume that the result

presentation of CPM studies should more frequently be perceived as understandable, useful, and trustworthy. For RQ2c and RQ3c, the binarized categories were 'difficult' (i.e., very difficult, difficult responses) and 'less difficult' (i.e., neutral, easy, very easy responses). This binarization was implemented to better capture the perception of significant difficulties that researchers face in authoring and reviewing understandable, useful, and trustworthy results in CPM studies. Regression analyses aimed to evaluate the significance of the associations between biomedical researchers' perceptions of study results in three quality attributes—understandable, useful, and trustworthy—and their demographic characteristics and experiences as authors and reviewers.

Qualitative data analysis

For the analysis of the qualitative data, the focus is to answer RQ4 of this study. I used thematic analysis (Braun and Clarke, 2006). Initially, I immersed myself in the data to gain a deep understanding of each answer provided by respondents. I excluded respondents' answers to the open-ended question about their reasoning regarding difficulties in producing understandable, useful and trustworthy CPM study results as authors and reviewers, if their responses were non-existent, have incomplete ideas, or fall outside the scope of challenges in authoring or reviewing. After excluding those answers, I generated themes to identify significant issues emerging from the data. Each theme should contribute to the overall narrative of the analysis. In the final phase, I ensured that each theme clarified the reasons behind the answers provided by respondents.

Ethics approval

I notified respondents of the consent process via recruitment emails. I informed respondents that they could proceed to the survey only if they provided informed consent to participate in the study by clicking on the link provided in the recruitment email. The University of Washington's Institutional Review Board (IRB) in Seattle determined this study to be exempt.

4.4 RESULTS

Survey responses and respondent characteristics

As detailed in Chapter 3, the study collected a total of 263 responses through two recruitment strategies: mailing lists (n=12) and PubMed email (n=251). This total includes all responses, even those where respondents did not click the 'submit' button at the end of the survey or have incomplete status (i.e., none of the questions in this survey were mandatory. Table 4.1 shows the survey completion status and the types of respondents. The majority of the surveys were complete. The table presents respondent types, either author, reviewer, or both author and reviewer. The number of publications determined the classification into these roles and reviews respondents reported, allowing for respondents to identify as both author and reviewer roles. The majority of respondents were both authors and reviewers.

Table 4.1 Survey and respondent characteristics.

Basic characteristics	n (%)
Survey responses (from both the mailing list and PubMed)	N = 263
Complete survey	221 (84%)
Incomplete survey	42 (16%)
Respondent types	N = 263
Author only	79 (30%)
Both author and reviewer	169 (64%)
Reviewer only	3 (1%)
Declined to state respondent type*	12 (5%)

* Excluded from analysis in RQ 3

Table 4.2 summarizes the demographic characteristics of the respondents. Most respondents fell within the age range of 35 to 65 years and identified as men who assume roles as both authors and reviewers. Additionally, most were affiliated with academic institutions and were primarily located in Europe and North America.

Table 4.2 Respondent characteristics

Characteristics	n (%)
Age	N = 263
18 - 24 years old	2 (1%)
25 - 34 years old	46 (17%)
35 - 44 years old	81 (31%)
45 - 65 years old	82 (31%)
66 - 79 years old	7 (3%)
Prefer not to say	2 (1%)
Declined to state	43 (16%)
Gender	N = 263
Man	146 (56%)
Woman	67 (25%)
I prefer to describe	1 (<1%)
Prefer not to say	5 (2%)
Declined to state	44 (17%)
Organization	N = 263
Academia	180 (68%)
Academia & other	16 (6%)
Other (Government, Consultancy, Freelancing, Industry)	16 (6%)
Declined to state	51 (19%)
Region	N = 263
Africa	7 (3%)
Asia	27 (10%)
Australasia	2 (1%)
Europe	90 (34%)
North America	61 (23%)
South America	10 (4%)
Declined to state*	66 (25%)

* Excluded from analysis in RQ 3

This study identifies respondents who assume both roles, as shown in Table 4.2, as authors and reviewers, as indicated in Table 4.3, which is determined by the number of publications they authored and reviewed. As a reminder, the inclusion criteria for this study required respondents to have experience publishing or reviewing research papers on CPM studies. Table 4.3 shows more

detailed experiences of respondents either as authors or reviewers. As indicated, there is an overlap between authors and reviewers, with the total number exceeding 263.

Table 4.3 Author and reviewer experience.

The term 'author' refers to respondents identified as 'author only' and as 'both author and reviewer' in Table 4.1. The term 'reviewer' is applied similarly.

Characteristics	Author n (%)	Reviewer n (%)
Published/ Reviewed studies	N = 248	N = 172
1 - 5	156 (63%)	89 (52%)
6 - 10	43 (18%)	37 (22%)
11 - 20	26 (10%)	18 (10%)
More than 20	23 (9%)	28 (16%)
Experience in year	N= 248	N= 172
Less than a year	5 (2%)	9 (5%)
1 to 5 years	117 (47%)	80 (47%)
6 to 10 years	68 (27%)	35 (20%)
11 to 20 years	42 (17%)	42 (24%)
More than 20 years	14 (6%)	5 (3%)
Declined to state	2 (1%)	1 (1%)
Types of published/ reviewed studies (preimpact analysis)	N= 248	N= 172
Developing and validating models	137 (57%)	110 (65%)
Developing models	83 (34%)	52 (31%)
Validating models	11 (3%)	3 (1%)
Declined to state	17 (6%)	7 (3%)
Types of published/ reviewed other than preimpact analysis studies	N= 67	N = 35
Impact Analysis	1(2%)	0
Methods	8(12%)	4(11%)
Prediction Model Development/Validation for Non-Clinical Purposes	12(18%)	4(11%)
Review	6(8%)	2(7%)
Systematic Review/Meta-analysis	20(30%)	12(34%)
Undefined	20(30%)	13(37%)

* Excluded from analysis in RQ 3

Also, from Table 4.3, both authors and reviewers are mostly in the early stages of their careers, as indicated by their participation in a limited number of studies and having up to 5 years

of experience. Their main research activity revolves around developing and validating models, with a significant number focusing solely on development. This points to a trend in which my respondents have more experience developing models than testing existing ones. Table 4.3 also highlights a minimal focus among the respondents on impact analysis studies—studies that assess the effectiveness of CPM in clinical settings.

4.4.1 *RQ1a: What is the frequency in which biomedical researchers perceive the results of CPM studies as understandable, useful, and trustworthy?*

Figure 4-1 illustrates the frequency with which respondents perceive the study results of CPM as understandable, useful, or trustworthy. The results for 'understandable' are slightly skewed towards the frequent categories ('always' and 'often'), indicating a tendency among respondents to find CPM results at least somewhat understandable. In contrast, 'useful' and 'trustworthy' show a skew towards the less frequent categories ('rarely' and 'never'), suggesting that respondents find CPM results less often useful or trustworthy.

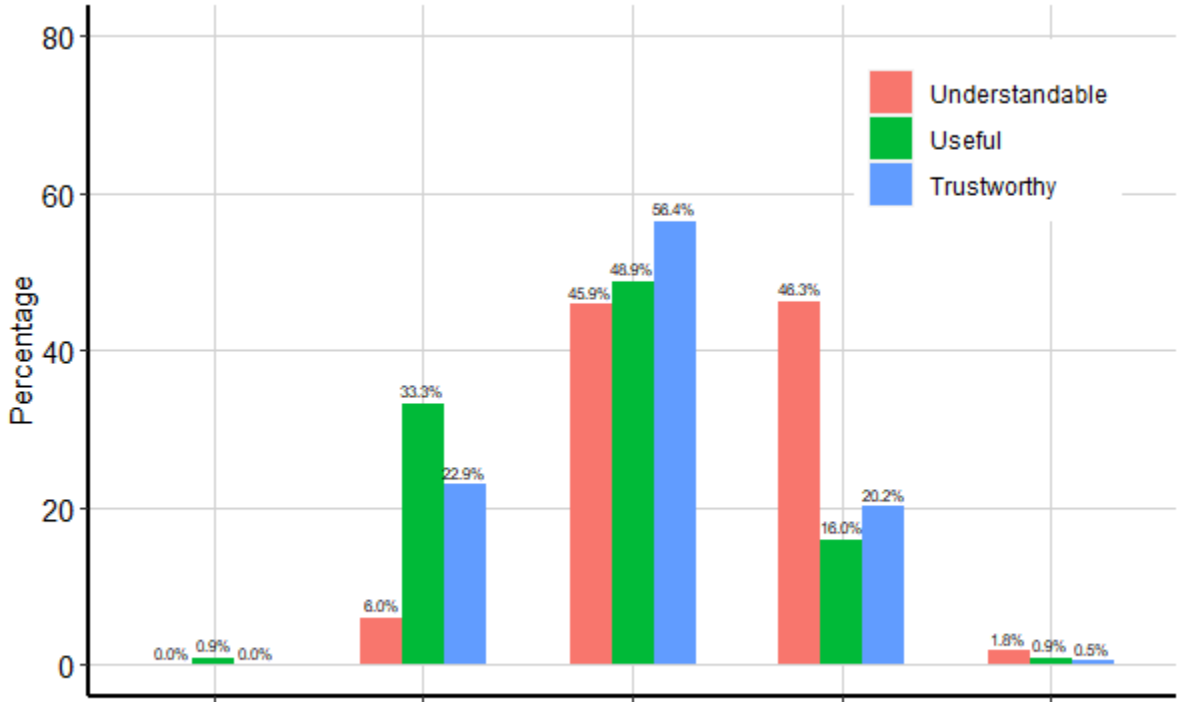

Figure 4-1 Distributions of respondents finding understandable, useful, and trustworthy study results in clinical prediction model research papers in (%) with N = 218

4.4.2 *RQ1b: Does the perception of biomedical researchers of CPM study results differ across the three quality attributes: understandable, useful, and trustworthy?*

The Friedman rank sum test results indicate a significant difference in the frequency at which respondents encounter understandable, useful, and trustworthy results in CPM studies ($F = 82$, $p<0.001$). This statistical evidence points to a differentiated pattern in how often respondents report each quality attribute—understandable, useful, and trustworthy—in CPM study results. A post-hoc pairwise comparison using the Nemenyi post-hoc test revealed significant differences. The frequency of perceiving useful results and understandable results differed significantly ($p < 0.001$). Similarly, the frequency of trustworthy results differed significantly from understandable results ($p = 0.002$). Also, the frequency of trustworthy results differed significantly from useful results ($p = 0.003$).

To support the Friedman rank sum test above, the boxplot in Figure 4-2 shows the distribution of respondents' ranks regarding finding results understandable, useful, and trustworthy in CPM studies. The y-axis represents rank with 1 as ‘never’ and 5 as ‘always.’ The median rank for 'understandable' is noticeably lower (median = 4.0) than for 'useful' and 'trustworthy' (both with a median of 3.0), indicating respondents typically find it less often to find useful results. The interquartile range (IQR) depicted in the box plot for each category shows 'useful' with the highest IQR, suggesting that respondents find results to be least frequently useful compared to 'understandable' and 'trustworthy.'

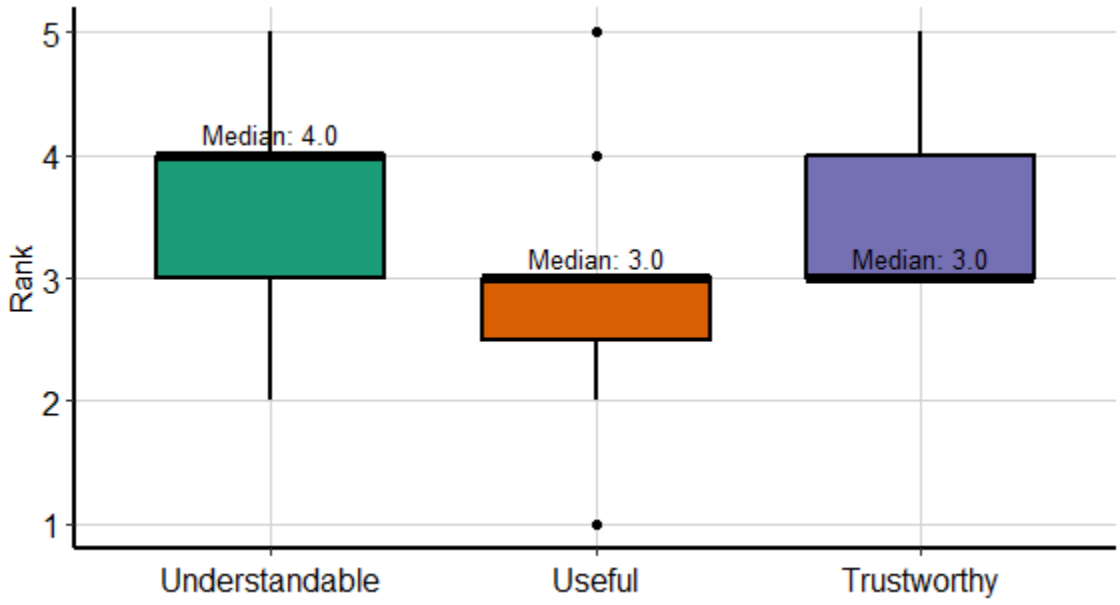

Figure 4-2 Distribution of rank for finding understandable, useful, and trustworthy results of CPM studies

4.4.3 *RQ1c: Do perceptions of CPM study results as understandable, useful, and trustworthy vary by biomedical researchers' characteristics?*

Table 4.4 displays the results of multivariable logistic regression evaluating the association between respondents' characteristics and their experiences with the frequency at which they find the results of CPM studies to be understandable, useful, and trustworthy.

Table 4.4 Respondents' characteristics associated with finding understandable, useful, trustworthy study results

	Respondents who frequently find study results. **understandable**			Respondents who frequently find study results **useful**			Respondents who frequently find study results **trustworthy**		
Characteristics	n (%)	OR [CI]	P-value	n (%)	Odds Ratio	P-value	n (%)	Odds Ratio	P-value
Demographic Characteristics									
Age									
18-34 years old	19(41.3)	Ref		10(21.7)	Ref		10(21.7)	Ref	
35-44 years old	38(48.1)	0.98[0.45-2.1]	0.9	19(23.8)	0.9[0.4-2.6]	0.98	15(18.8)	1.2[0.45-3.3]	0.68
45 years old & above	45(51.7)	0.63[0.28-1.4]	0.3	7(8.1)	2.4[0.8-7.5]	0.13	19(22.1)	0.8[0.29-2]	0.62
Gender									
Man	69(48.2)	Ref		23(16)			30(20.8)	Ref	
Woman	31(48.4)	1[0.52-1.9]	0.9	13(20.3)	0.9[0.4-2.1]	0.75	13(20.3)	1.19[0.5-2.7]	0.68
Region									
America	35(50.7)	Ref		7(10.1)	Ref		15(21.7)	Ref	
Europe	42(46.7)	1.3[0.68-2.5]	0.4	13(14.4)	0.7[0.3-1.9]	0.53	14(15.6)	1.7[0.7-3.9]	0.21
Other	15(45.5)	1.13[0.47-2.7]	0.8	13(38.2)	0.2[0.06-0.5]	0.0018	9(26.5)	0.75[0.3-2]	0.56
Respondent types									
Author or reviewer	30(54.5)	Ref		12(21.8)	Ref		15(27.3)	Ref	
Both	75(46)	1.64[0.82-3.3]	0.2	25(15.2)	1.4[0.5-3.3]	0.50	30(18.4)	2.01[0.9-4.5]	0.09
Author characteristics									
Research papers authored									
1-5 publication	67(48.2)	Ref		31(22.1)	Ref		30(21.4)	Ref	
6-10 publication	20(55.6)	0.9[0.4-1.9]	0.7	2(5.6)	3.4[0.9-22.4]	0.12	7(20)	0.76[0.3-2.2]	0.59
More than 10	15(37.5)	2.29[0.9-5.9]	0.07	4(10)	1.42[0.4-5.9]	0.6	7(17.5)	0.7[0.2-2.2]	0.54

	Respondents who frequently find study results. **understandable**			Respondents who frequently find study results **useful**			Respondents who frequently find study results **trustworthy**		
Author experience year									
5y or less	48(44.4)	Ref		25(23.1)	Ref		24(22.2)	Ref	
6 to 10y	28(50.9)	0.6[0.3-1.3]	0.2	8(14.3)	1.51[0.59-4.3]	0.41	14(25)	0.9[0.4-2.3]	0.92
More than 10y	26(50)	0.4[0.2-1.1]	0.1	4(7.7)	2.14[0.6-9.3]	0.27	6(11.8)	2.3[0.7-8.2]	0.16
Study types of research papers authored									
Develop & validate	51(41.8)	Ref		15(12.2)	Ref		22(18)	Ref	
Develop/ validate	43(53.8)	0.5[0.3-0.9]	0.04	20(25)	0.57[0.26-1.2]	0.16	17(21.2)	0.9[0.4-1.9]	0.75
Reviewer characteristics									
Research papers reviewed									
1-5 publications	42(50)	Ref		15(17.6)	Ref		19(22.6)	Ref	
6-10 publications	16(44.4)	1.3[0.54-3.3]	0.5	7(19.4)	0.52[0.17-1.6]	0.24	7(19.4)	1.1[0.4-3.6]	0.85
More than 10	20(43.5)	1.6[0.57-4.7]	0.4	3(6.5)	0.89[0.2-4.9]	0.89	5(10.9)	1.5[0.4-7.1]	0.55
Reviewer experience years									
5y or less	42(50)	Ref		19(22.4)	Ref		21(24.7)	Ref	
6 to 10y	12(35.3)	1.41[0.58-3.5]	0.5	5(14.7)	1.9[0.62-6.8]	0.29	5(14.7)	1.7[0.6-6]	0.35
More than 10y	24(51.1)	0.7[0.25-1.9]	0.5	1(2.1)	14.19[2.1-290.1]	0.02	5(10.9)	2.6[0.6-13]	0.2
Study types of research papers reviewed									
Develop & validate	47(44.8)	Ref		16(15.1)	Ref		21(19.8)	Ref	
Develop/ validate	28(51.9)	0.86[0.43-1.7]	0.7	9(16.7)	1.02[0.4-2.7]	0.96	9(17)	1.5[0.64-4]	0.34

For 'understandable,' no significant associations were found with demographic characteristics (i.e., age, gender, region) and respondent type (i.e., author, reviewer, or both). Among authors, frequently finding results understandable was associated with experience in writing research papers that develop or validate models; these respondents were more frequently to find results understandable (OR: 0.53, CI: [0.28-0.98], p=0.04) compared to those involved in both types of studies. No significant associations were observed for other authors' experiences, including the number of papers authored and the number of years of experience as authors. Among reviewers, no significant associations were found with their experience (i.e., the number of papers reviewed, years as reviewers, and types of studies reviewed) in frequently finding CPM results understandable.

For 'useful,' the only demographic characteristic demonstrating a significant association with frequently finding results 'useful' was the respondents' region of origin. Those from regions other than America and Europe were more frequent to find the results useful (OR: 0.18, CI: [0.058-0.51], p< 0.01). There were no significant associations with other demographic characteristics or respondent types. The only authors' experience showing a significant association was the number of years as an author. Those with more than 10 years (OR: 14.19, CI: [2.1-290.1], P-value: 0.02) found study results useful less frequently compared to those with 5 years or less. Other experiences as an author did not have significant associations. For reviewers, no experience, including the number of publications, years of experience as reviewers, or types of studies reviewed, showed significant associations with frequently finding results that were useful.

For 'trustworthy,' no significant associations were identified for any characteristics or experiences, either as authors or reviewers. However, it is important to note that all findings with $p < 0.05$ presented relatively wide confidence intervals.

4.4.4 *RQ2a: What is the level of difficulty that biomedical researchers as authors experience when producing understandable, useful, and trustworthy results in CPM studies?*

Figure 4-3 depicts the level of difficulty that authors encounter in producing study results that are understandable, useful, and trustworthy. These authors also include those who have experience as reviewers as well. The chart shows that authors' experiences in producing understandable results of CPM studies are predominantly clustered in the less difficult categories ('very easy,' 'easy,' and 'neutral'). Conversely, 'useful' and 'trustworthy' responses are considerably categorized as 'difficult' and 'very difficult,' indicating that authors often encounter challenges in ensuring CPM study results are useful or trustworthy.

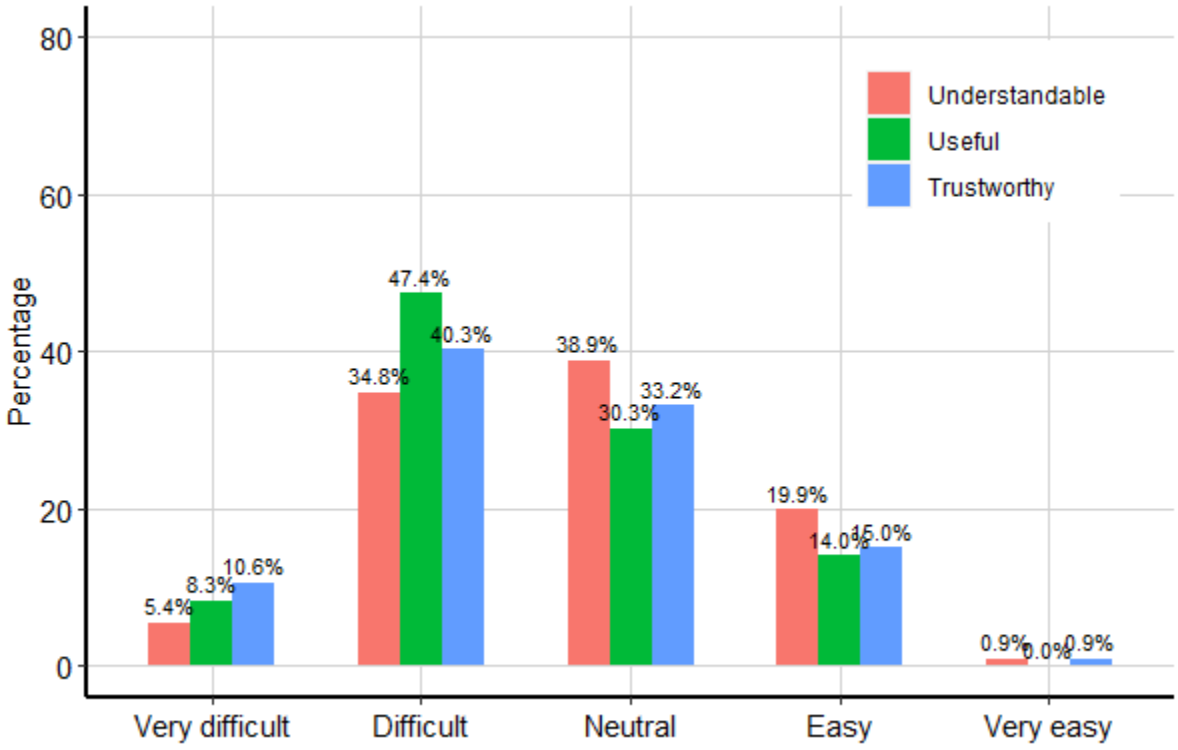

Figure 4-3 Distributions of authors experiencing difficulty in producing understandable, useful, and trustworthy study results (in %) with N = 218

4.4.5 *RQ2b: Do the level of difficulties experienced by biomedical researchers as authors in producing understandable, useful, and trustworthy results of CPM studies differ across the three quality attributes?*

The results from the Friedman rank sum test suggest significant differences in the difficulties experienced by authors in producing understandable, useful, and trustworthy results in CPM studies ($F=20$, $p<0.001$). At least one of the quality attributes—understandable, useful, or trustworthy—poses a different level of challenge for authors when generating results that contain these quality attributes. Nemenyi post-hoc tests found differences in pairwise comparisons of these quality attributes. The analysis identified a statistically significant difference in difficulty between useful results and understandable results ($p = 0.009$). The difference in difficulty for producing 'trustworthy' versus 'understandable' results was also significant ($p = 0.043$). On the other hand, the test shows no difference in difficulty level for 'useful' and 'trustworthy' results.

The boxplots in Figure 4-4 illustrate these statistical findings by illustrating the distribution of ranks for each attribute. The y-axis represents a scale from 1 for 'very difficult' to 5 for 'very easy'. Both 'useful' and 'trustworthy' results share a higher median difficulty rank of 2.0, which signifies a greater difficulty than 'understandable' results in CPM studies, which have a lower median rank of 3.

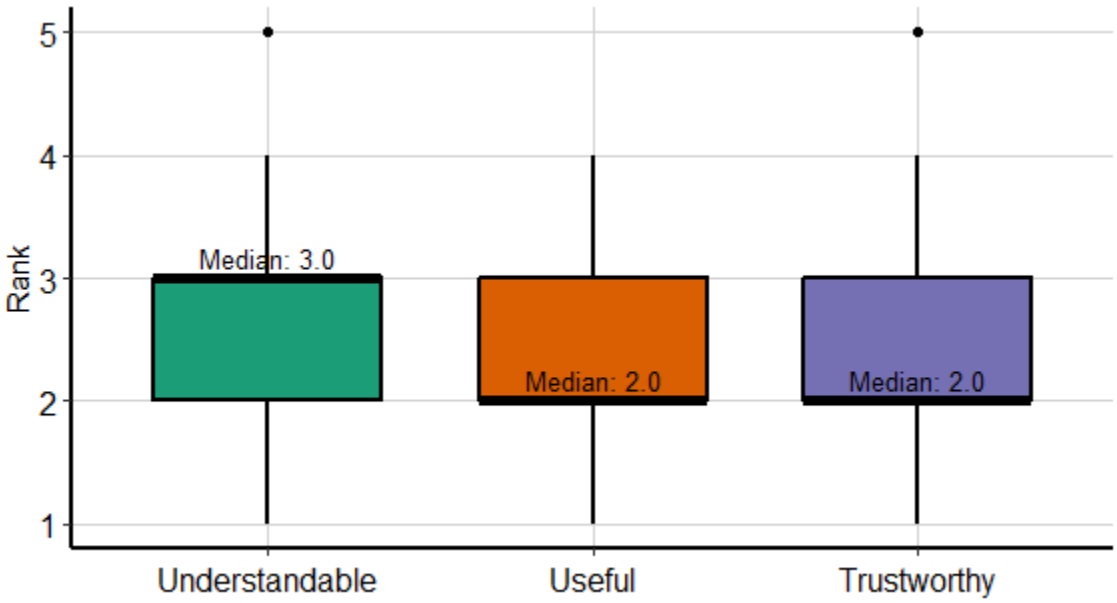

Figure 4-4 Distribution of rank for the difficulties experienced by authors to produce understandable, useful, and trustworthy results of CPM studies

4.4.6 *RQ2c: Do experiences in encountering difficulties in authoring understandable, useful, and trustworthy results vary by biomedical researchers' characteristics?*

Table 4.5 presents results from a multivariable logistic regression that assesses whether experiences of encountering difficulties among respondents as authors in ensuring understandable, useful, and trustworthy results vary by their characteristics and experiences. The outcomes for difficulty are categorized into 'difficult' (encompassing 'very difficult' and 'difficult') and 'less difficult' (including 'neutral,' 'easy,' and 'very easy').

Table 4.5 Author characteristics and their experience in authoring study results

	Authors who experience difficulties in authoring **understandable** study results			Authors who experience difficulties in authoring **useful** study results			Authors who experience difficulties in authoring **trustworthy** study results		
Author characteristics	n (%)	Odds Ratio	P-value	n (%)	Odds Ratio	P-value	n (%)	Odds Ratio	P-value
Age									
18-34 y.o	19(42)	Ref		27(59)	Ref		30(65)	Ref	
35-44 y.o.	36(47)	0.96[0.4-2.3]	0.93	51(65)	0.54[0.22-1.3]	0.18	45(57)	1.21[0.5-3]	0.68
45 y.o. & above	27(32)	1.57[0.59-4.2]	0.36	40(47)	1.27[0.48-3.4]	0.63	33(39)	2.58[0.97-7.1]	0.06
Gender									
Man	54(39)	Ref		82(58)	Ref		79(56)	Ref	
Woman	24(40)	0.72[0.35-1.5]	0.37	31(48)	1.41[0.7-2.9]	0.33	26(41)	1.91[0.95-3.9]	0.07
Region									
America	30(44)	Ref		39(57)			32(47)	Ref	
Europe	32(38)	1.42[0.68-3]	0.35	54(61)	0.76[0.36-1.6]	0.47	51(57)	0.72[0.34-1.5]	0.37
Other	13(39)	1.46[0.57-3.8]	0.43	17(50)	1.27[0.5-3.2]	0.62	20(59)	0.86[0.33-2.2]	0.76
Research papers authored									
1-5 publ	61(42)	Ref		80(54)	Ref		81(55)	Ref	
6-10 publ	12(34)	0.72[0.26-2]	0.54	21(57)	0.54[0.19-1.5]	0.23	17(46)	0.83[0.3-2.2]	0.70
More than 10	16(40)	0.36[0.11-1.1]	0.08	26(60)	0.36[0.12-1]	0.06	17(40)	1.25[0.44-3.5]	0.67
Author experience year									
5y or less	53(47)	Ref		71(62)	Ref		67(59)	Ref	
6 to 10y	21(38)	2.21[0.93-5.5]	0.08	29(48)	1.99[0.83-4.8]	0.12	30(52)	0.79[0.33-1.9]	0.59
More than 10y	15(28)	3.98[1.2-14.5]	0.03	27(50)	2.92[0.92-9.7]	0.07	18(33)	1.58[0.52-4.9]	0.42
Study types of research papers authored									

	Authors who experience difficulties in authoring **understandable** study results			Authors who experience difficulties in authoring **useful** study results			Authors who experience difficulties in authoring **trustworthy** study results		
Develop & validate	48(38)	Ref		73(57)	Ref		56(44)	Ref	
Develop/ validate	39(47)	0.91[0.45-1.8]	0.79	50(58)	1.19[0.6-2.4]	0.61	53(62)	0.59[0.29-1.2]	0.14
Respondent types									
Author only	27 (47)	Ref		37(62)	Ref		36(62)	Ref	
Both author and reviewer	62(38)	0.81[0.35-1.8]	0.6	90(54)	1.08[0.4-2.5]	0.86	79(47)	1.13[0.5-2.6]	0.77

For 'understandable,' the only significant association is between those who have more than 10 years of experience as an author who experiences less difficulty in producing understandable results, compared to those with 5 years or less (OR: 3.98, CI: [1.2-14.5], p=0.029). There were no significant associations with other author characteristics (i.e., age, gender, region) and experiences (i.e., number of publications and types of studies authored). Furthermore, no significant associations existed between 'useful' and 'trustworthy' with either author characteristics or experiences. It is also important to note that the findings with P-values < 0.05 presented relatively wide confidence intervals.

4.4.7 *RQ3a: What is the level of difficulties that biomedical researchers as reviewers experience difficulties when providing peer review feedback to ensure understandable, useful, and trustworthy results in CPM studies?*

Figure 4-5 displays the level of difficulty reviewers encounter in providing feedback to achieve understandable, useful, and trustworthy study results. These reviewers also include those who have experience as authors as well. The chart suggests that reviewers perceive providing feedback to ensure understandable results of CPM studies as leaning towards the less difficult categories ('very easy,' 'easy,' and 'neutral'). However, for 'useful' and 'trustworthy,' there is a notable shift toward the 'difficult' categories ('difficult,' 'very difficult'), indicating reviewers' experience more difficulty in providing feedback to ensure CPM study results are 'useful' or 'trustworthy.'

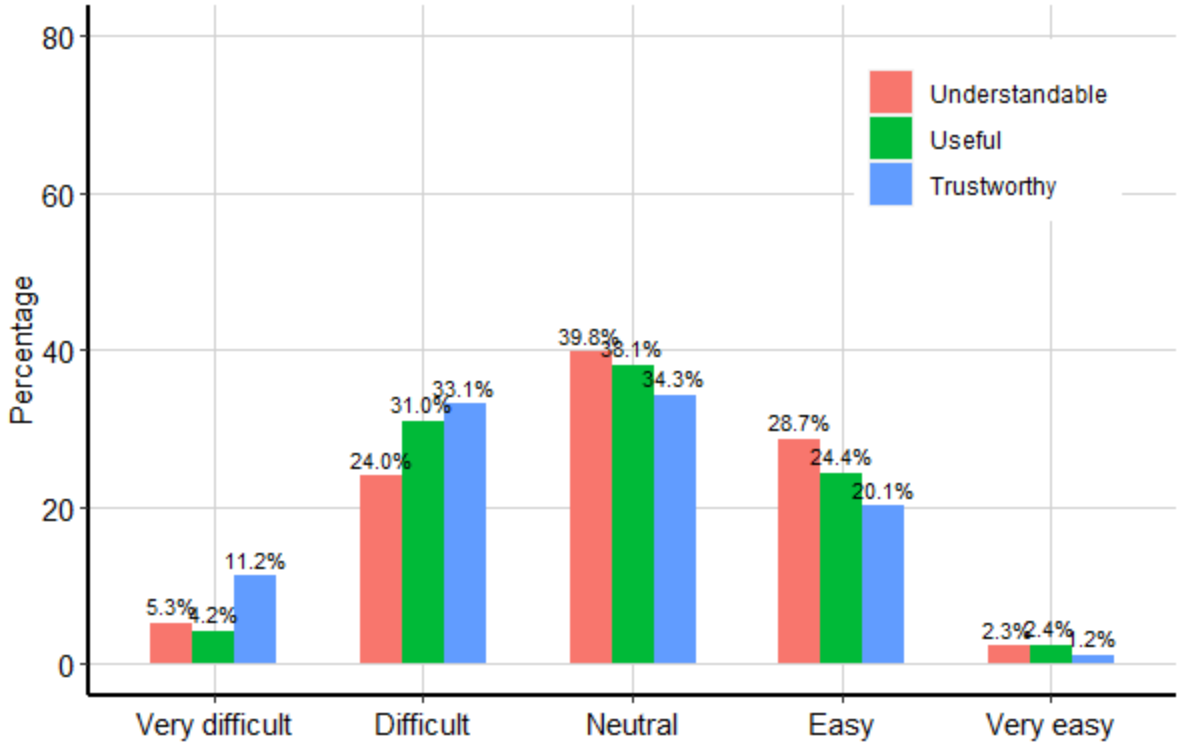

Figure 4-5 Distributions of reviewers experiencing difficulties in providing feedback for understandable, useful, and trustworthy study results (in %) with N = 218

4.4.8 *RQ3b: Do the level of difficulties experienced by biomedical researchers as reviewers in providing feedback to ensure understandable, useful, and trustworthy results of CPM studies differ across the three quality attributes?*

The results of the Friedman test indicate that reviewers perceive different levels of difficulties when providing feedback to ensure understandable, useful, and trustworthy results of CPM studies ($F(2) = 29$, $p < 0.001$), highlighting the distinct experiences of reviewers across these three quality attributes. Nemenyi post-hoc comparisons found the challenge in providing feedback for trustworthy results compared to understandable results significantly differs ($p = 0.004$). Furthermore, the difference in difficulty between 'useful' and 'trustworthy' was noted to be trending

towards significance (p = 0.07). On the other hand, the test showed no difference between 'understandable' and 'useful.'

The boxplots in Figure 4-6 illustrate the statistical findings, offering a graphical representation of the reviewers' difficulty ranks for each quality attribute—understandable, useful, and trustworthy results of CPM studies. The y-axis represents a scale from 1 for 'very difficult' to 5 for 'very easy'. All three criteria have a median rank of 3.0, indicating that, on average, reviewers find them moderately challenging. While 'understandable' and 'useful' share similar interquartile ranges, implying no significant difficulty difference, 'trustworthy' stands out with a wider range and an outlier, hinting at greater difficulty than ‘understandable’ as shown in the Nemenyi post-hoc comparison. Moreover, the marginal significance between 'useful' and 'trustworthy' suggests a trend that 'trustworthy' may also be perceived as slightly more challenging than 'useful.'

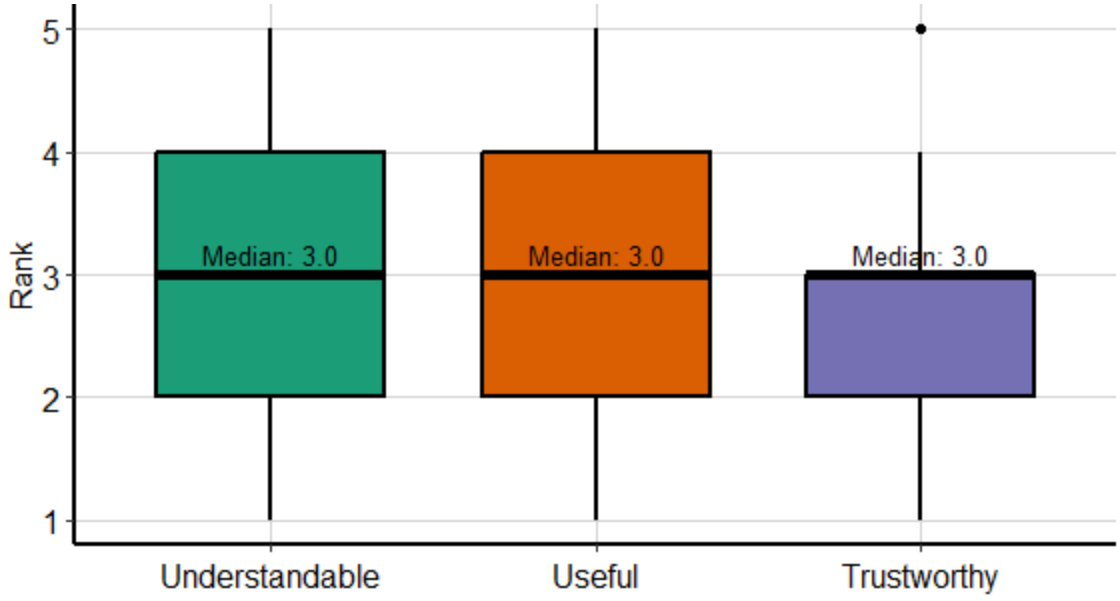

Figure 4-6 Distribution of rank for the difficulties experienced by reviewers to provide feedback to ensure understandable, useful, and trustworthy results of CPM studies

4.4.9 *RQ3c: Do experiences encountering difficulties in reviewing understandable, useful, and trustworthy results vary by biomedical researchers' characteristics?*

Table 4.6 presents results from multivariable logistic regression regarding the difficulties reviewers encountered when providing peer review feedback to ensure that results are understandable, useful, and trustworthy, segmented by reviewer characteristics and experience. Similar to the analysis for authors, the outcomes for difficulties are categorized into 'difficult' (encompassing 'very difficult' and 'difficult') and 'less difficult' (including 'neutral,' 'easy', and 'very easy').

Table 4.6 Reviewer characteristics and their experience in authoring study results

	Reviewers who experience difficulties in providing feedback to ensure **understandable** study results			Reviewers who experience difficulties in providing feedback to ensure **useful** study results			Reviewers who experience difficulties in providing feedback to ensure **trustworthy** study results		
Reviewer characteristics	n (%)	Odds Ratio	P-value	n (%)	Odds Ratio	P-value	n (%)	Odds Ratio	P-value
Age									
18-34 y.o.	11(39)	Ref		13(46)	Ref		16(57)	Ref	
35-44 y.o.	25(40)	1.21[0.42-3.5]	0.72	24(39)	1.99[0.7-5.74]	0.19	34(54)	1.46[0.51-4.25]	0.48
45 y.o. & above	14(19)	4.59[1.33-16.8]	0.02	22(31)	2.3[0.74-7.41]	0.16	24(33)	4.23[1.32-14.44]	0.03
Gender									
Man	31(26)			37(32)	Ref		49(42)	Ref	
Woman	16(37)	0.37[0.15-0.9]	0.05	20(48)	0.4[0.17-0.92]	0.03	23(53)	0.33[0.13-0.78]	0.01
Region									
America	14(26)			20(38)			25(47)		
Europe	20(30)	0.87[0.35-2.1]	0.76	25(38)	1.16[0.51-2.65]	0.73	29(44)	1.29[0.57-2.94]	0.54
Other	12(46)	0.35[0.08-0.9]	0.05	9(35)	1.28[0.43-3.98]	0.66	14(54)	0.79[0.26-2.36]	0.67
Research papers reviewed									
1-5 publ	27(31)			31(36)	Ref		40(47)	Ref	
6-10 publ	13(35)	0.92[0.31-2.8]	0.89	14(38)	1.02[0.37-2.93]	0.97	17(46)	1.32[0.47-3.84]	0.59
More than 10	10(22)	1.78[0.45-7.9]	0.43	14(31)	0.95[0.27-3.4]	0.94	18(39)	2.38[0.68-9.09]	0.19
Reviewer experience year									
5y or less	29(33)			30(35)	Ref		39(45)	Ref	
6 to 10y	11(31)	0.49[0.15-1.5]	0.22	16(46)	0.37[0.13-1.06]	0.07	17(50)	0.33[0.1-0.99]	0.05
More than 10y	9(19)	0.53[0.1-2.6]	0.43	12(26)	1.03[0.25-4.18]	0.97	18(38)	0.32[0.07-1.3]	0.12

	Reviewers who experience difficulties in providing feedback to ensure **understandable** study results			Reviewers who experience difficulties in providing feedback to ensure **useful** study results			Reviewers who experience difficulties in providing feedback to ensure **trustworthy** study results		
Study types of research papers reviewed									
Develop & validate	33(30)	Ref		36(33)	Ref		49(45)	Ref	
Develop/ validate	17(31)	1.36[0.56-3.4]	0.51	23(43)	0.71[0.31-1.61]	0.41	24(44)	1.09[0.48-2.52]	0.84

For 'understandable,' three reviewer characteristics were associated with more difficulty in providing feedback to ensure understandable results of CPM studies. Reviewers aged 18 – 34 have more difficulties providing feedback that ensures understandable results (OR: 4.59, CI: [1.33-16.8], P-value: 0.018) compared to those aged 45 and above. Women and those from regions other than America and Europe found it more difficult to provide feedback that ensures understandable results compared to men and those from America, with (OR: 0.37, CI: [0.15-0.9], p=0.029), and (OR: 0.35, CI: [0.08-0.9], p= 0.046) respectively.

For 'useful,' the only significant association was gender. Women found it more difficult to provide feedback that ensures useful results (OR: 0.4, CI: [0.17-0.92], p = 0.032).

For 'trustworthy,' reviewers aged 18 – 34 have more difficulties in providing feedback that ensures trustworthy results compared to those aged 45 and above (OR: 4.23, CI: [1.32-14.44], p= 0.017). Women perceived it as more difficult to provide feedback compared to men (OR: 0.33, CI: [0.13-0.78], p= 0.014). No other significant associations were found between other characteristics and experiences as reviewers. However, the wide confidence intervals (CIs) indicate a higher level of uncertainty around the point estimates, suggesting that while the results are statistically significant, caution is advised in their interpretation.

4.4.10 *RQ4: What are the reasons underlying challenges that respondents experience?*

My qualitative analysis, based on 88 open-ended responses, identified six explanatory themes concerning the challenges of producing CPM research papers. It is worth noting that not all responses were included in the analysis; exclusions were made for responses from 175 respondents due to their not answering, incomplete ideas, or those falling outside the scope of challenges in authoring or reviewing to ensure understandable, useful, and trustworthy results in CPM studies. Examples of these excluded answers are in Appendix D. Table 4.7 shows the

distribution of respondents whose responses to the open-ended questions are included in the qualitative analysis.

Table 4.7 Distribution of respondents whose responses to the open-ended questions included or not included in the analysis

Characteristics/ Experiences	**Included n (%)**	**Not Included n (%)**
Age	N = 88	N = 175
18-34 y.o.	24(27.3)	24(13.7)
35-44 y.o.	25(28.4)	56(32)
45 y.o. & above	37(42)	52(29.7)
Declined to state	2(2.3)	43(24.6)
Gender	N = 88	N = 175
Man	63(71.6)	83(47.4)
Woman	20(22.7)	47(26.9)
Declined to state	5(5.7)	45(25.7)
Organization	N = 88	N = 175
Academia	68(77.3)	112(64)
Academia & other	7(8)	9(5.1)
Other (Government, Consultancy, Freelancing, Industry)	7(8)	9(5.1)
Declined to state	6(6.8)	45(25.7)
Region	N = 88	N = 175
Africa	4(4.5)	3(1.7)
Asia	6(6.8)	21(12)
Australasia	2(2.3)	0(0)
Europe	35(39.8)	55(31.4)
North America	29(33)	32(18.3)
South America	4(4.5)	6(3.4)
Declined to state	8(9.1)	58(33.1)
Respondent types	N = 88	N = 175
Author or reviewer	22(25)	60(34.3)
Both author and reviewer	66(75)	103(58.9)
Declined to state	0(0)	12(6.9)

Characteristics/ Experiences	Included n (%)	Not Included n (%)
Published studies	N = 88	N = 175
1-5 publication	59(67)	97(55.4)
6-10 publication	12(13.6)	31(17.7)
More than 10	17(19.3)	32(18.3)
Declined to state	0(0)	15(8.6)
Reviewed studies	N = 88	N = 175
1-5 publication	33(37.5)	56(32)
6-10 publication	16(18.2)	21(12)
More than 10	17(19.3)	29(16.6)
Declined to state	22(25)	69(39.4)
Author experience	N = 88	N = 175
5y or less	44(50)	78(44.6)
6 to 10y	26(29.6)	42(24)
more than 10y	18(20.4)	38(21.7)
Declined to state	0(0)	17(9.7)
Study type (Author)	N = 88	N = 175
Develop & validate	48(54.5)	89(50.9)
Develop/ validate	36(40.9)	58(33.1)
Declined to state	4(4.5)	28(16)
Reviewer experience	N = 88	N = 175
5y or less	35(39.8)	54(30.9)
6 to 10y	15(17.1)	20(11.4)
More than 10y	16(18.2)	31(17.7)
Declined to state	22(25)	70(40)
Study type (Reviewer)	N = 88	N = 175
Develop & validate	44(50)	66(37.7)
Develop/ validate	21(23.9)	34(19.4)
Declined to state	23(26.1)	75(42.9)

From the interview analysis, I identified three groups of themes related to difficulties in Group 1) presenting understandable, useful, and trustworthy CPM study results (Theme 1 – 3),

Group 2) conducting CPM studies (Theme 4 - 5), and Group 3) providing reviews to authors (Theme 6). These complete themes include Difficulties in 1) understanding CPM results among the target users, 2) presenting clinically relevant and useful prediction models, 3) ensuring trustworthy results of CPM studies, 4) acquiring and utilizing quality data, 5) addressing complex and evolving methodologies in predictive model studies, and 6) providing feedback to authors among reviewers. In the next section, I describe each theme with representative quotes. Throughout the description of each theme, I indicate which respondent IDs contributed to each theme and select the most relevant quote accompanied by their IDs.

The list of selected quotes is provided in Appendix E, presenting the original text from the interviews. The selection of these quotes for each theme is based on those that initially exhibit similar ideas representative of each sentence within the theme.

Theme 1: Difficulties in understanding clinical prediction models among the target users

Respondents identified challenges in understanding CPMs, focusing on the complex interplay between model complexity and user background (IDs 7, 63, 93, 94, 107, 117, 127, 143, 198, 231, 240, 260). The difficulties in conveying complex statistical concepts to a broader audience, specifically healthcare providers, underlie these challenges (IDs 7, 93, 107, 143, 198, 236, 239, 240, 260). Specific statistical and methodological knowledge, which may not be present in the target audience, further exacerbates this issue (IDs 93, 107, 198, 231). Additionally, tension exists between the level of statistical detail that a prediction model should contain and what an average clinical reader typically comprehends (IDs 9, 73). Many respondents affirmed this inherent complexity in making prediction models understandable (IDs 9, 68, 93, 198, 229). One respondent captured this theme, stating:

" *It is a challenge to describe statistical analyses and results of a prediction model paper in understandable language for the target audience, which typically includes health care providers. Understanding the results and implications of prediction models does require specific statistical and methodological knowledge."* (ID 198)

Theme 2: Difficulties in presenting clinically relevant and useful prediction models

Presenting clinically relevant and useful CPM study results poses a complex challenge, especially when considering the complexities of clinical practices in which the target settings of CPM studies are situated (IDs 22, 26, 79, 84, 87, 111, 127, 148, 220, 254). Respondents noted that the development of CPM requires alignment with real-world healthcare needs and prioritizes clinical applicability (IDs 66, 74, 82, 221). Furthermore, the quest for presenting the clinical relevance of CPMs goes beyond model accuracy, necessitating a model that resonates with different populations and translates into actionable insights for clinical practices (IDs 42, 210, 249). One respondent encapsulated this complexity, stating:

'Well, it is one thing to build a predictive model that's significant using a cutting-edge methodology. But it is a totally different ballgame if one wishes for these results to be useful, meaningful, and reliable to clinicians". (ID 41)

Theme 3: Difficulties in ensuring trustworthy clinical prediction models

Respondents identified complex issues in ensuring trustworthy results of CPM studies. Among reviewers, the main issues ranged from the lack of accessibility to original data, and codes, to difficulties in replicating the study pipeline, and the models that authors included in their manuscripts (IDs 73, 75, 88, 236, 242). For reviewers, reporting those parts in the study results reflects the that authors were being honest about what they really did in the study, which are the main components of 'trustworthy' (IDs 15, 29, 30, 33, 36). Among authors, the complexities extend

to the evolving methodologies and data quality factors (IDs 66, 75, 107, 107, 242). Reporting these factors accurately has become the main concern in ensuring that their study results are trustworthy (IDs 15, 29, 33, 36, 88). For instance, one respondent noted the challenge of ensuring trustworthy reporting due to the "black box" nature of the models that they used, which also contain inherent sources of bias (IDs 29, 63, 107). Another respondent emphasized the difficulty in making models trustworthy, which involves rigorous data quality and methodology checks in their manuscripts (IDs 15, 29, 36, 107). One respondent succinctly summarized these challenges:

> "*One of the key challenges is trustworthiness because of poor knowledge of the situations where prediction models miss, black box models, history of coding in inherent sources of bias, and inherent distrust of automation*." (ID 107)

Theme 4: Difficulties in acquiring and utilizing quality data

Another challenge described by both authors and reviewers is related to difficulties encountered when conducting a CPM study. This challenge is specifically about obtaining high-quality data for predictive modeling, particularly due to its limited availability (IDs 35, 106, 163, 172, 179, 204, 259), lack of access to comprehensive databases, and constraints related to data sharing in -healthcare (IDs 86, 106, 163, 204, 208). Even when data is available, operational issues that affect data quality, such as missing data, small sample sizes, and a lack of standardized clinical measures, can hinder the effectiveness and reliability of the models (IDs 10, 19, 20, 106, 163, 165, 171, 172, 184, 187, 198, 204, 208). These factors collectively contribute to concerns about the reliability and replicability of predictive models, affecting their utility in various research and clinical settings (IDs 10, 20, 106, 117, 163, 172, 196, 198, 204, 259). As one respondent aptly noted:

"*Due to restrictions on clinical data access and further use/dissemination, it is extremely difficult to produce useful and reliable study results. Many studies focus on a specific region or even an organization, limiting the variability of the data used and even the number of attributes available to develop the models.*" (ID 106)

Theme 5: Difficulties in navigating complexity and evolving methodologies conducting studies in clinical prediction models

The complexity and challenges in conducting a CPM study are often multifaceted and often underestimated, creating a landscape that is difficult to navigate for both researchers and practitioners (IDs 58, 118, 120, 153, 179). One significant issue is the variability in methodologies used to construct these models, coupled with the rapidly evolving and sometimes opaque methods that make it difficult to keep pace with best practices (IDs 8, 58, 63, 95, 118, 153, 179, 182, 204, 216). Additionally, the rise of machine learning models and AI use has introduced a new layer of complexity; these advanced algorithms are often non-interpretable (IDs 63, 204, 229). As one respondent pointed out:

"*The AI studies I have reviewed are more difficult because there is less conceptual rigor to the ideas, and they often seem remote from practical clinical interpretation...*" (ID 229)

Theme 6: Difficulties in providing feedback to authors among reviewers

Reviewing to ensure understandable, useful, and trustworthy CPM studies are often perceived as less challenging than authoring (IDs 62, 79, 85, 95, 107, 117, 158, 163, 196, 233, 260). Fewer challenges may also arise from their own authorial experiences (IDs 79, 92, 233). Specific challenges in providing reviews include ensuring consistency in validation and replication. Respondents raise specific concerns that challenges in providing reviews are particularly evident with authors from fields other than biostatistics or clinical epidemiology, who

might display deficits in foundational knowledge about prediction models, potentially leading to pitfalls (IDs 12, 62, 85, 117, 163, 196). Another challenge emerges when ensuring a model's utility, especially when its value deviates from standard interpretations (IDs 3, 10, 12, 85, 158). While some reviewers perceive the act of giving feedback as straightforward, its actual incorporation by authors to enhance the model's clinical relevance can be more challenging (IDs 74, 85, 158, 196). As one respondent reported,

> "*It is easy to provide feedback, but it does not mean it is easy 1) to ensure they are completely relevant given the clinical field; 2) to have the feedback accepted by the authors*." (ID 158)

4.5 DISCUSSION

Overall, my study's quantitative findings indicate challenges prevalent across all three quality attributes (understandable, useful, trustworthy) in perceiving (RQ1a), authoring (RQ2a), and reviewing (RQ3a) CPM study results. When perceiving the quality of CPM study results, 'useful' emerges as the most challenging quality attribute, followed by 'trustworthy' and 'understandable' (RQ1b). Authors find demonstrating 'useful' and 'trustworthy' results more challenging compared to 'understandable' results (RQ2b). Reviewers consider 'trustworthy' results more challenging to achieve than 'understandable' results, but equally challenging as 'useful' results (RQ3b). Despite some respondents' characteristics showing significant associations with the challenges of perceiving (type of paper, region, reviewer experience years), authoring (reviewer experience years), and reviewing (age, gender, region) CPM study results that meet the three quality attributes, many other characteristics are not associated, suggesting widespread challenges among biomedical researchers overall (RQ1c), including authors (RQ2c) and reviewers (RQ3c). In further qualitative analysis, participants acknowledge the difficulties in presenting CPM study

results as understandable, useful, and trustworthy (RQ4), highlighting challenges such as difficulties in understanding clinical prediction models among target users, presenting clinically relevant and useful prediction models, and ensuring trustworthy CPM study results.

Perceptions of biomedical researchers about the quality of CPM study results

For biomedical researchers, ‘useful’ emerges as the most challenging quality attribute, followed by 'trustworthy', when they are asked about their perception of CPM study results in research papers around the three quality attributes. The significant differences are between the frequency of perceiving 'useful' results and 'understandable' results, as well as between 'trustworthy' results and both 'understandable' and 'useful' results. However, the median indicates that 'useful' results are perceived less frequently compared to both 'understandable' and 'trustworthy' results. This could be due to their lack of involvement in the study's internal processes, which may lead to skepticism or uncertainty about whether the results are useful and trustworthy. Without firsthand knowledge of how the study was developed to align with clinical practices, the use of study methodologies, data collection, and analysis procedures, they may find it more difficult to justify whether the study results are useful and trustworthy. On the other hand, perceiving 'understandable' results of CPM research papers as the least challenging may be due to the fact that all respondents are biomedical researchers with experience as authors of CPMs, meaning they are relatively knowledgeable and skilled in the field to understand the content of the CPM study results.

Factors such as demographic and personal characteristics, including experience and socioeconomic status, are often utilized to explain perceptions, practices, and outcomes related to the three quality attributes—understandable, useful, and trustworthy—embedded in information sources through association analysis. For example, in the context of 'understandable,' rooted in

Bloom's Taxonomy, a study investigating whether demographic and socioeconomic factors are associated with learning outcomes found significant associations (Morgan, 2011). Regarding 'useful,' as conceptualized within the TAM, another study demonstrated that demographic factors are influential in shaping attitudes and practices toward using online information (Porter and Donthu, 2006). In the case of 'trustworthy,' a study focusing on demographics and personal characteristics, such as political affiliations, revealed their impact on trust in news sources (Verma et al., 2018). Unlike these studies, which primarily target lay populations, including students and the general public, my study consists predominantly of domain experts in a specific field, CPM, evidenced by their experiences as authors and reviewers of CPM research papers.

My study's findings suggest that certain factors are associated with perceiving CPM study results as understandable and useful but not trustworthy. Notably, authors involved in both developing and validating CPM studies tend to perceive results as more understandable, possibly due to their exposure to different types of CPM studies, which may help them find the results more comprehensible. Interestingly, reviewers with less than 5 years of experience perceive CPM study results as more useful compared to those with over 10 years of experience. This difference could be attributed to reviewers with more extended experience, who have encountered a greater number of CPM research papers over time, which do not align well with the clinical realities, as mentioned in the systematic review studies introduced in prior work in Chapter 1. Additionally, authors from outside America and Europe are more likely to find CPM studies useful, possibly because they find CPM more applicable to their clinical practice.

However, other potentially important factors, such as the number of publications and experiences as reviewers, were not found to be associated with perceiving CPM study results as understandable, useful, or trustworthy. Even though certain factors influence the perception of

CPM study results as understandable and useful, the lack of association with most other demographic and experiential factors among respondents suggests that challenges in perceiving CPM results as understandable, useful, and trustworthy are widespread across various respondent demographics and experiences as authors and reviewers. Additionally, the target population of this study can be categorized as domain experts with extensive expertise in the CPM field. Yet, the percentages of those who always or often perceive CPM study results as understandable, useful, and trustworthy are not that high compared to those who perceive them as sometimes, rarely, or never. This may indicate that the root of the problem lies not within the target populations but in the inherent qualities of the CPM results being presented.

Challenges as authors

As authors, the challenge reported by biomedical researchers shifts towards demonstrating that both 'useful' and 'trustworthy' have comparable challenges and are more challenging compared to 'understandable.' As experienced authors, it seems that reporting the studies as understandable is less challenging because it mostly involves detailing what they have done in the studies. With 'useful,' they must ensure that the results align with the clinical practice problems they aim to address, which should have been developed well before implementing the studies and motivating the research. Achieving this can be difficult when many unforeseen events occur during the study that do not align with the initial motivations, making the reporting of useful results problematic. For instance, in a CPM study focused on developing early disease detection, biomarkers initially identified as potential predictors from the literature may later be found to have been poorly collected in clinical practice, casting doubt on whether the model would be useful. For 'trustworthy,' authors may find it challenging to report the study results because they need to

balance revealing enough information to demonstrate that their study is of high quality and worthy of publication while also being transparent about the study's potential limitations.

Studies focusing on identifying the challenges faced by authors—or producers—of information sources in ensuring that information meets the three quality attributes of understandable, useful, and trustworthy are still not as common as those examining challenges faced by readers of this information. With limited information available in this area, my study sought to determine whether any respondent characteristics or experiences as authors are associated with challenges in producing CPM study results that embody these qualities. The logistic regression analysis revealed a significant association only for authors with more than 10 years of experience, who reported fewer difficulties in producing understandable results compared to those with 5 years or less, possibly due to increased confidence in presenting CPM results gained through experience. However, the fact that other factors did not show associations, combined with a significant portion of authors—about 40% for 'understandable' and more than 50% for 'useful' and 'trustworthy'—reporting difficulties, suggests that a substantial number of authors experience challenges in meeting these quality attributes in their work.

Challenges as reviewers

As reviewers, respondents consider providing feedback to ensure that CPM study results 'understandable' to be as challenging as ensuring they are 'useful.' This may be because, as those who were not directly involved in the studies being reported, achieving useful CPM results that align with clinical practices can be as challenging as clearly presenting them to ensure the audience understands. They may not really know how clinical practice truly informs and motivates the study. This might drive skepticism about whether the study results are understandable and useful to them. For 'trustworthy,' reviewers consider it more challenging to achieve than 'understandable'

yet equally as challenging as 'useful,' with a caveat: the p-value for transitioning from 'useful' to 'trustworthy' is 0.07, close to the standard threshold of 0.05. These results suggest that reviewers may exhibit heightened skepticism when trying to assess if the study results were reported in a trustworthy manner or as they truly happened. This heightened skepticism may stem from reviewers' concerns that authors might report results not as they are but in a manner that increases the likelihood of acceptance for publication, thereby influencing the perception that the reported study results are trustworthy (George, 2020).

Similar to authors, reviewers are also producers of information, where studies identifying the challenges, they face in ensuring information meets the quality attributes of understandable, useful, and trustworthy are also less common than those focused on the readers or users of the information. My analysis found that age, gender, and region were the only factors significantly associated with challenges in providing feedback to ensure CPM results meet these quality attributes. Specifically, younger reviewers aged 18–34 reported more difficulties in providing feedback for understandable and trustworthy results compared to those aged 45 and above, possibly due to early career challenges and the absence of clear guidance in providing the feedback to ensure that CPM study results meet the three quality attributes. Additionally, reviewers outside North America and Europe faced more challenges in providing feedback and ensuring understandable CPM study results, potentially due to language barriers, given the predominance of English in academic journals. However, these challenges were not consistent across all quality attributes. A notable finding was that female reviewers consistently reported more difficulties across all three quality attributes. A potential reason is gender differences in self-confidence in providing critical feedback (Lenney, 1977), highlighting a need for further exploration into the peer review process dynamics. Nevertheless, since most other factors were not associated with

challenges in providing feedback to ensure CPM study results meet the three quality attributes, my study suggests that these challenges are widespread across respondent demographic characteristics and experiences.

Other factors that could explain study findings

Another factor that may explain my study's findings relates to the concepts of 'understandable,' 'useful,' and 'trustworthy' themselves. Perceiving and ensuring 'understandable' results is consistently regarded as the least challenging aspect compared to 'trustworthy' among respondents. Except among reviewers, ensuring 'understandable' results is also considered the least challenging aspect compared to 'useful.' This phenomenon may stem from the fact that 'understandable' seems to be a more familiar concept compared to the other two within academic communities, which form the background of most biomedical researchers who participated in this study.

The origin of 'understandable' can be traced back to Bloom's Taxonomy of Educational Objectives, published in 1956 in the field of education (Bloom, 1956). In education, where the presentation and assimilation of information are integral and daily activities, information inherently becomes a core component of the teaching and learning process. More specifically, subsequent developments have expanded the use of the cognitive domain in Bloom's Taxonomy to measure information understanding among students during the learning process (Lord and Baviskar, 2007). Furthermore, in biomedical informatics, the quality attribute of 'understandable' is used to measure the quality of presented information, particularly in data visualization (Burns et al., 2020; Mahyar et al., 2015). Given the established foundation of 'understandable' within Bloom's Taxonomy, rooted in the educational field where respondents have accumulated extensive experience throughout their learning journeys, respondents seem to be more confident in

identifying and delivering understandable results. In contrast, 'useful' and 'trustworthy' appear to be relatively newer concepts in this context, as will be further discussed in the following paragraph.

The concept of 'useful' can be traced back to the perceived usefulness in management information systems for accounting, introduced by Larcker and Lessig in the early 1980s (Larcker and Lessig, 2007). Subsequently, Fred D. Davis published a study in 1989 focusing on user acceptance of information technology alongside the concept of ease of use (Davis, 1989). There, the usefulness centered around the use of electronic mail, later known as email. Subsequent work by others, including Venkatesh, popularized 'usefulness' as part of the Technology Acceptance Model (TAM) (Venkatesh and Bala, 2008), which expanded the concept's application to measure usefulness across various information technology domains and beyond. For presenting information beyond information technology, I could not find relevant literature utilizing this concept of 'useful' as extensively as 'understandable,' which has become a common quality to achieve in presenting information across a broad spectrum, such as in education and biomedical informatics. This might explain why biomedical researchers are less accustomed to focusing on 'useful' as compared to 'understandable,' which potentially leads to lower confidence in identifying and delivering useful results, especially among respondents when perceiving the quality of CPM study results and producing CPM study results as authors.

For 'trustworthy,' the earliest literature combining this concept with presenting information is about trust in digital information, published by Kelton in 2008 (Kelton et al., 2008). Earlier literature that discussed the trust concept focused on psychological concepts and did not integrate with information (Corazzini, 1977; Gambetta, 2000). The literature focusing on similar topics around trust and health information, such as Jadad et al., did not provide a conceptual framework that explains trustworthiness in information like Kelton et al. did (Jadad and Gagliardi, 1998; Lampe

et al., 2003). Kelton proposed an integrated model of trust in information, detailing how information sources possessing trustworthy attributes such as accuracy, objectivity, validity, and stability would support acting upon the information for decision-making. Later applications of this concept include identifying the trustworthiness of various information sources, such as those from information systems and libraries (Donaldson, 2016; Meeßen et al., 2020). Similar to 'useful,' the concept of trust in presenting information, such as study results of a research project, is relatively more recent compared to 'understandable.' This recency may contribute to biomedical researchers being less accustomed to identifying and delivering 'trustworthy' results.

Reasons for difficulties experienced by authors and reviewers

The findings of RQ4 provide context on the difficulties encountered by authors and reviewers in ensuring that the results of CPM studies yield understandable, useful, and trustworthy results. When answering the question about the reasons why respondents perceive the level of difficulty in ensuring CPM study results are understandable, useful, and trustworthy when authoring or reviewing, I identified three groups of themes. Group 1 is related to themes about difficulties when presenting the study results, Group 2 is related to themes about difficulties encountered during the conduct of CPM studies, and Group 3 is related to difficulties specific to reviewers.

Theme group 1 highlights the shared challenge among respondents of making the results of CPM studies understandable, useful, and trustworthy. For 'understandable,' they emphasize the need for enhanced communication when presenting CPM study results, acknowledging the inherent difficulty of conveying complex information to a broader, less specialized audience such as clinicians as the target users of CPM. Efforts should, therefore, focus on making the results as accessible as possible to these audiences.

For 'useful,' respondents emphasized that CPM studies should be clinically relevant by aligning them with the practical demands of healthcare settings. This means that CPM studies should reflect careful planning and design by the authors from the beginning of the study. For example, if a study is initially poorly designed or lacks clinical relevance, producing useful results is nearly impossible. Having data alone is not adequate for developing robust models; such data must be grounded in clinical realities, often requiring collaboration among multiple stakeholders. A lack of clinical relevance complicates the task of demonstrating a useful result. Even if authors are able to explain their findings clearly and make them understandable to a broad audience, making the results useful may still be questioned without this critical clinical connection. Thus, when authors fail to design their studies with clinical practices in mind, for example, by just relying on the availability of data to develop CPM, respondents might find these models lack grounding and do not address challenges in clinical practices.

For 'trustworthy,' respondents highlight the need for addressing barriers to trust to CPM study results such as the limited presentation of results, restricted access to original data and codes for developing or validating models, and the application of responsible methodologies. Therefore, achieving trustworthy results seems to require that authors present their studies comprehensively, revealing all aspects. Comprehensive reporting could lead to the discovery of inconsistencies, potentially compromising the efforts to ensure trustworthy results. These factors may contribute to the greater challenge of ensuring trustworthy results in CPM studies compared to making the results understandable.

In addition to the theme group that directly provides context to the challenges encountered by authors and reviewers in ensuring understandable, useful, and trustworthy results of CPM studies, my study also identified two additional theme groups. In Theme Group 2, Themes 4 and

5 pinpoint the foundational challenges when conducting CPM studies. Acquiring and effectively utilizing quality data remains a consistent challenge, with issues ranging from access restrictions to concerns over data variability (Theme 4). Furthermore, another challenge involves keeping up with rapidly advancing techniques, especially with the advent of machine learning and AI, which demands continuous adaptation and learning from researchers (Theme 5). These two challenges seem to contribute to respondents' reasoning when discussing the difficulty level in presenting CPM study results. Thus, according to respondents, presenting understandable, useful, and trustworthy results involves processes such as selecting methodologies and ensuring data quality from the beginning of CPM studies.

The last Theme Group I identified focuses on difficulties expressed specifically by reviewers—Theme 6. This theme offers a perspective among reviewers that centers on their primary concerns when providing feedback to authors. Although reviewers may find it relatively easy to give feedback, integrating this feedback to enhance the clinical relevance and utility of CPMs is crucial and often not straightforward for authors to incorporate. This insight highlights the significance of open dialogue between authors and reviewers to ensure that CPM studies' insights are rigorous and clinically relevant. This theme underscores that the interaction between reviewers and authors also contributes to the reasoning behind respondents' answers as reviewers.

Thus, respondents clearly acknowledge the challenges in presenting CPM study results as understandable, useful, and trustworthy. The thematic analysis reveals that communicating results to a broader audience, especially clinicians, ensuring clinical relevance, and maintaining transparency in methodologies and results are pivotal in overcoming these obstacles. This acknowledgment underscores the necessity for further exploration and targeted efforts to address

these issues. This exploration may involve further clarifying these challenges and identifying items that need to be reported in CPM study results to address those challenges.

Implications to my study findings

Measuring attitudes toward the three quality attributes—understandable, useful, and trustworthy—like in this study, is an important concept in psychology. In psychology, an attitude is a tendency expressed by evaluating a particular entity with some degree of favor or disfavor (Verplanken and Orbell, 2022). Prior studies often show that attitudes toward particular topics, such as attitudes toward risk factors of a particular disease, differ according to respondents' demographic characteristics, such as age, gender, and education. Then, the results indicating whether each characteristic is more favored compared to others will inform follow-up actions or recommendations, such as tailoring messages or education for those characteristics (Jensen et al., 2012). These measurements often depend on the distribution of the respondents' characteristics that are suitable for such analysis, indicating whether one characteristic will show differences in attitudes over another.

My data shows variations in the distribution of characteristics, such as age, gender, and types of experiences, allowing me to examine whether there are certain characteristics associated with their attitudes toward perceiving, authoring, or reviewing to ensure CPM study results are understandable, useful, and trustworthy. The findings could inform a focus on select characteristics if any significant associations are shown. Unfortunately, the significant associations are not quite consistent across the three roles of respondents when perceiving the quality of CPM study results and producing CPM study results as authors and reviewers. This means that respondents' attitudes are likely not influenced by their demographics or experiences, such as the number of papers published or reviewed and years of involvement in the field of CPMs. The challenges are dispersed

among these groups. Thus, follow-ups for this study, such as identifying needs or solutions or tailoring messages to ensure understandable, useful, and trustworthy results in reporting CPM studies, may target broader biomedical researchers who do not necessarily focus on specific demographic characteristics and experience.

Furthermore, the respondents of this survey are predominantly individuals with experience in preimpact analysis CPM studies. The qualitative portion of the study results, particularly in Themes 1 through 5, indicates that most challenges referenced by respondents pertain to preimpact analysis CPM studies, which involve the development and validation of CPMs. Given this context, addressing the challenges identified by these respondents should specifically focus on improving the quality of CPM study results to be understandable, useful, and trustworthy, with a particular emphasis on preimpact analysis studies.

Limitations of the Study

This study provides valuable insights into the challenges faced by authors and reviewers in producing understandable, useful, and trustworthy results in CPM research papers. However, several limitations should be noted:

1. The subjects of this research are biomedical researchers, mostly newer researchers as indicated by relatively few publications but with sufficient experience, as evidenced by their history of publishing or reviewing studies in the CPM field. This study excluded new researchers who have not yet published in this field, which may limit the generalizability of the findings to all types of biomedical researchers. However, this approach also ensures that the judgments of the respondents, when answering questions for RQ1, were not hindered by a lack of knowledge and skills in implementing CPM studies, which might have been the case had the study involved a less experienced target population.

2. In this study, most respondents assumed two different roles, as authors, and reviewers. A potential concern may arise since our respondents often overlap, with 70% being both authors and reviewers. While a logistic regression analysis conducted in this study found no significant differences in the challenges faced by those exclusively as authors and those both as authors and reviewers in authoring understandable, useful, and trustworthy results of CPM studies (RQ2C), future research could benefit from distinctly analyzing three separate groups: those who are exclusively authors, those who are both authors and reviewers and those who are only reviewers. However, such a distinction was not feasible in this study due to the very low number of respondents who were solely reviewers (only 3 respondents).
3. The use of a Likert scale in this study incorporated middle options such as 'sometimes' and 'neutral,' which resulted in unexpectedly high percentages for these choices. This approach follows common practice, as suggested by references (Brown, 2010). To address the issues arising from these middle options, I employed the Friedman test to examine differences in ratings. I found significant differences when analyzing respondents' challenges, whether when perceiving the quality of CPM study results (RQ1b),and producing CPM study results as authors (RQ2b), or reviewers (RQ3b). These significant results from the Friedman test bolstered my confidence that the responses, even with higher percentages in the middle options, are meaningful.
4. Another potential limitation of using the Likert scale is losing context regarding the reasoning behind respondents' answers since it is limited to the options provided within the 5-point scale. My qualitative follow-up somewhat mitigated these limitations. The questionnaire offered open-ended questions allowing authors and reviewers to explain the

reasoning behind their selections, as reported in the RQ4 result. A few insights were shared about why respondents selected 'neutral,' but these were not consolidated into a single theme in the RQ4 results, as they did not directly answer the research questions. From these few responses, I got the impression that respondents selected 'neutral' because they acknowledged difficulties in reporting or providing feedback on the study results but felt confident enough to address these challenges.

5. Limitations of this study may also come from the type of recruitment, as described in Chapter 3, that determines who is involved in this survey. Biomedical researchers recruited may be limited to those captured from PubMed records with specific keyword searches used in the search strategies. Biomedical researchers whose publications in CPM do not fall within those search strategies or are listed in databases outside PubMed, such as CINAHL or EMBASE, may not be recruited. Furthermore, as this study captured relatively new researchers, future research, in addition to expanding keyword searches, should not only use PubMed but also other databases and may also limit recruitment to those with more experienced researchers. These limitations may also apply to Chapters 5 and 6, as the participants' recruitment used the same flow as described in Chapter 3
6. Using English for the survey while targeting multiple countries could have introduced bias by excluding non-English speakers or those less proficient, potentially skewing results towards English-dominant regions or individuals.

These limitations should be considered when interpreting our findings.

Future studies

This study appears to be the first to assess challenges experienced by biomedical researchers related to perceiving, authoring, and reviewing understandable, useful, and trustworthy

results. Although RQ4 attempted to identify specific challenges, the findings seem limited and do not provide detailed information on the needs of biomedical researchers to overcome these challenges. An interview study that identifies the need will be reported in Chapter 5.

Future research should also pivot towards clinicians to understand their perspectives on CPM study results. These studies could explore the specific challenges clinicians face in perceiving CPM study outcomes as understandable, useful, and trustworthy. Direct engagement with clinicians can provide valuable insights into those expected to adopt CPMs in clinical practice and tailor the reporting of CPM study results to meet their needs. This focus on clinicians could lead to the development of CPMs that are easy to use and improve diagnostic and prognostic accuracy, as highlighted in my survey with clinicians in Chapter 2, ultimately enhancing the efficacy of clinical decision-making processes.

4.6 Conclusions

This study quantitatively demonstrated challenges experienced by biomedical researchers across three quality attributes—understandable, useful, and trustworthy—of CPM study results among biomedical researchers, whether as authors, or reviewers. These challenges are prevalent across multiple demographic characteristics and experiences of biomedical researchers, indicating that the challenges are pervasive among biomedical researchers. The inclusion of qualitative analysis deepens my understanding, unveiling the complex challenges researchers face in ensuring CPM results meet these three quality attributes. These challenges involve difficulties in presenting understandable, useful, and trustworthy CPM study results, conducting CPM studies, and providing reviews to authors. Thus, identifying these challenges is crucial for determining whether the challenges of ensuring the quality of CPM study results across the three quality attributes

present among biomedical researchers whether as authors or reviewers of CPM study results underscores the need to address these challenges to enhance the quality of CPM study results.

Chapter 5. CHARACTERIZING THE NEEDS OF BIOMEDICAL RESEARCHERS FOR UNDERSTANDABLE, USEFUL, AND TRUSTWORTHY RESULTS IN CLINICAL PREDICTION MODEL STUDIES: AN INTERVIEW STUDY

5.1 INTRODUCTION

Biomedical researchers assert that the quality of CPM research papers, specifically for preimpact analysis studies, should be understandable, useful, and trustworthy (Damen et al., 2016; Dhiman et al., 2021; Heus et al., 2018; Yusuf et al., 2020). However, the main reporting guidelines for these studies, TRIPOD, which offers suggestions for improving the quality of CPM research papers, focus more on transparency quality attributes than on these three quality attributes. Guidelines similar to TRIPOD (Moons et al., 2015), including PROBAST (Wolff et al., 2019), and Guidelines for Developing and Reporting Machine Learning Predictive Models in Biomedical Research (Luo et al., 2016) also lack a focus on these three quality attributes.

Chapter 4 of my dissertation revealed the challenges of perceiving and presenting understandable, useful, and trustworthy results among biomedical researchers as authors or reviewers. Further quantitative analysis from the survey also reveals that these three quality attributes have very different challenges in presenting CPM study results. Specifically, as authors, biomedical researchers encounter significantly greater difficulty in producing 'useful' and 'trustworthy' results compared to 'understandable' results. As reviewers, producing trustworthy results seems to be significantly more challenging compared to producing understandable results.

In the qualitative portion of Chapter 4, biomedical researchers articulated the challenges of ensuring the quality of CPM study results across these three quality attributes, suggesting a need for further exploration. Drawing upon human-centered design principles, identified challenges in a system or product, in this case, the report of CPM study results should be addressed by involving its target users, which in my study are biomedical researchers, to understand the challenges further and identify requirements or needs to address those challenges (FDIs, 2009; Shrier et al., 2020). Therefore, more robust qualitative studies with comprehensive data collection methods, such as interviews, may offer deeper insights into biomedical researchers' needs to ensure understandable, useful, and trustworthy results of CPM studies—perspectives that are currently lacking in the well-known guidelines for reporting CPM studies.

Understandable, useful, and trustworthy as information qualities of reporting study results

Advocating for reporting results of CPM studies, in which this study refers to the reporting in peer-reviewed research papers, to be understandable, useful, and trustworthy aligns with information qualities suggested in previous studies. As introduced in Chapter 1, these three qualities find their basis in Wang and Strong's 1996 framework for data quality, which was later adapted to information quality (Ge, 2009; Wang and Strong, 1996). Furthermore, the terms ‘understandable,’ ‘useful,’ and ‘trustworthy’ also have roots in descriptions that can be closely related to assessing the quality of information in research papers. ‘Understandable’ is associated with users' perceptions, as reflected in concepts like the six levels of understanding in Bloom's taxonomy (Lord and Baviskar, 2007; Mahyar et al., 2015). ‘Useful’ is frequently linked to perceived usefulness in the Technology Acceptance Model (TAM) (Larcker and Lessig, 2007; Venkatesh, 2000). ‘Trustworthy’ is commonly used in discussions about trust in digital information (Kelton et al., 2008; McGuinness and Leggatt, 2006; Meeßen et al., 2020).

Unfortunately, these three characteristics of information quality—understandable, useful, and trustworthy—are not commonly incorporated into standard practices or reporting guidelines such as TRIPOD and PROBAST for quality attributes in presenting the results of CPM studies.

Qualitative research in biomedical informatics

Qualitative research in biomedical informatics may involve: 1) Understanding needs, which explores user requirements, captures workflows, and comprehends users' mental models to inform the development of systems that align with those needs; 2) Design, which focuses on the system development process; 3) Evaluation, which examines how people interact with the system and assesses the impact of deployed systems in real settings; 4) Conceptual advancement, which involves defining the nomenclature, policies, curricula, and research spaces within biomedical informatics (Hussain et al., 2020). Therefore, understanding the need for understandable, useful, and trustworthy results of CPM studies represents an area of qualitative research within the biomedical informatics domain. Furthermore, using deductive approaches in qualitative studies allows for research about understanding needs based on existing concepts from prior research for further description (Ancker et al., 2021; Hsieh and Shannon, 2005). Since the three quality attributes—understandable, useful, and trustworthy—have been the subjects of extensive previous research regarding information quality, employing a deductive approach enables the leveraging of this substantial body of work to identify and meet the needs of biomedical researchers for results of CPM studies that embody these quality attributes.

5.2 STUDY OBJECTIVE

In my survey study (Chapter 4), I reported that biomedical researchers encountered significant challenges in authoring and providing peer review that ensures the study results of CPM research papers are understandable, useful, and trustworthy. In this follow-up interview study with

a sample of survey respondents, my objective was to characterize the needs of biomedical researchers to ensure understandable, useful, and trustworthy study results in CPM research papers. To achieve this objective, I answer the following research question (RQ): *What needs do biomedical researchers express for CPM study results to be understandable, useful, and trustworthy?*

5.3 METHODS

Study Design

This study employs a qualitative study design, utilizing individual remote interviews with authors and reviewers of CPM research papers. Interviews allow me to tap into needs by asking participants about their experiences of authoring and reviewing papers.

Recruitment

Participants were selected from survey respondents (Chapter 4) who opted to be contacted for a 60-minute follow-up interview. The survey recruitment process and inclusion criteria are described in Chapter 4. Upon indicating their interest in the follow-up interview, respondents were provided a link to review consent information and agree to participate, provide their email address, and choose a convenient time for the remote interview. The remote setup allowed for flexibility in scheduling and ensured that participants from various locations could easily contribute to the study.

Data collection

The data collection was conducted through 60-minute remote semi-structured interviews in English, focusing on identifying biomedical researchers' needs for ensuring quality results of CPM studies in three quality attributes: understandable, useful, and trustworthy. The interviews began with a brief introduction to the study, obtaining verbal consent, and posing questions related

to each quality attribute, allowing participants to respond in three dedicated “time periods.” Before asking questions about each quality attribute, I reminded participants of its definition. The interview questions covered the needs of biomedical researchers to ensure understandable, useful and trustworthy CPM study results, and their visualization preferences in supporting CPM study results that also meet the three quality attributes. This approach was informed by challenges identified in Chapter 4 and included questions about visualization preferences inspired by literature on using visualizations to enhance the presentation quality of information and achieve prediction models that meet the three quality attributes. The literature basis for including visualization preferences and the findings regarding visualizations are presented in Chapter 6. The interview guide can be found in Appendix F.

Of the 263 survey respondents, 50 confirmed their willingness to participate in the interview. I reached out to all 50 of these individuals to arrange interviews, adhering to a protocol of contacting each participant up to three times to confirm their final consent and schedule the interviews. Of these, thirteen did not respond to scheduling emails, and one email was undeliverable due to an incorrect email address. Additionally, 12 participants responded once but did not engage in further follow-up, while four participants did not show up for the scheduled interviews and reschedules were not set up. In total, I conducted interviews with 20 participants. Among these, two were treated as "trial" interviews, and the remaining 18 were included in the analysis. The trial interviews led to adjustments in the interview questions, optimizing them for clarity and relevance.

The interviews were all conducted remotely, primarily conducted using Zoom, with some utilizing Microsoft Teams. Most interviews were completed within 45 minutes, with some lasting one hour and others approximately 30 minutes. The duration of the interviews was carefully

regulated, ensuring that no interview was shorter than 30 minutes or exceeded 1 hour. This time frame of the interviews was established to provide sufficient depth while maintaining a focused and manageable conversation. Each interview proceeded through three distinct phases to discuss the three theme categories: ‘understandable,’ ‘useful,’ and ‘trustworthy’ study results. I recorded and transcribed interviews for qualitative analysis.

Data analysis

Transcripts were analyzed qualitatively using directed content analysis, aligning with the deductive approach of qualitative study design (Hsieh and Shannon, 2005). Such an approach allows for the utilization of a priori concepts of interest or relationships among them, assisting in forming the initial coding scheme and the interrelations between codes.

In my directed content analysis, I utilized a codebook comprising three quality attributes of reporting results of CPM studies: understandable, useful and trustworthy. These quality attributes were drawn from the expectations of biomedical researchers (Damen et al., 2016; Dhiman et al., 2021; Heus et al., 2018; Yusuf et al., 2020) and in accordance with the following frameworks: Bloom's taxonomy (Burns et al., 2020) to define the quality characteristics of ‘understandable’ perceived usefulness as commonly identified in the Technology Acceptance Model (TAM) (Hyun et al., 2009) for ‘useful,’ and from concepts of trust found in Trust in Information Systems for ‘trustworthy’ (Kelton et al., 2008). By adopting these frameworks, I created a codebook containing three themes, ‘understandable,’ ‘useful,’ and ‘trustworthy’ along with their corresponding sub-themes. For ‘understandable,’ the theme consists of six sub-themes (i.e., or ‘codes’): knowledge, comprehension, application, analysis, synthesis, and evaluation. For ‘useful,’ the 8 sub-themes include critical support, easier to use, more accomplishment, quality increase, useful for tasks, performance increase, quicker to complete, and user-control. For

‘trustworthy,’ the 4 sub-themes included accuracy, objectivity, validity, and stability. All sub-themes mentioned above are derived from previous studies: the sub-theme for 'understandable' from Burns et al. (Burns et al., 2020), 'useful' from Hyun et al. (Hyun et al., 2009), and 'trustworthy' from Kelton et al. (Kelton et al., 2008).

Table 5.1 provides the original definitions for each sub-theme and my working definitions. The original definition of each sub-theme originates from previous studies. Conversely, the working definition was generated by contextualizing these definitions during the analysis of the interviews. The analysis adhered to distinct time periods for each theme (i.e., understandable, useful, and trustworthy); the analysis on ‘understandable’ codes was analyzed when participants addressed this theme, and the same approach was applied to ‘useful’ and ‘trustworthy.’

Table 5.1 Definitions of codes and themes used in the qualitative analysis

Theme	Sub-theme	Original definition	Working definition
Understandable (Burns et al., 2020)	Knowledge	Recall basic facts and definitions	Participants mention their need for the results to provide basic facts and definitions.
	Comprehension	Understand the information in context	Participants express their desire for results to report the information in context.
	Analysis	Break down a concept into parts and understand their relationship	Participants want the results to break down concepts into parts and understand their relationship.
	Application	Apply knowledge to a new problem or represent it differently	Participants discussed the need for results to demonstrate the application of knowledge from CPMs to new problems, clinical settings, or populations or represent the models in different ways.
	Synthesis	Use knowledge to create something new	Participants mentioned that the results should facilitate their desire to see the use of knowledge being presented as able to create or initiate something new, like giving new approaches to solving a particular clinical problem.

Theme	Sub-theme	Original definition	Working definition
	Evaluation	Judge the value of information backed by evidence	Participants want the results to help them judge the value of presented information backed by evidence.
Useful (Hyun et al., 2009)	Critical support	Support critical aspects of my required/ desired task	Participants expect the results to present how the prediction models would support critical elements of the end users' required or desired tasks; otherwise, the end users may fail at their tasks if the results don't deliver such presentations.
	Easier to use	Make it easier for me to use for my tasks/ work	Participants want the results to show that it will be easier for end users to use CPMs when performing their tasks or work.
	More accomplishment	Accomplish more tasks than would otherwise be possible	Participants discuss the need for the results to help users accomplish more tasks than would otherwise be possible.
	Quality increase	Increase the quality of my tasks/ work	Participants express the desire for results to show how CPMs would increase the quality of users' tasks or work.
	Useful for tasks	Find the study results useful for my tasks	Participants want the results to be useful for readers' tasks as researchers or clinicians.
	Performance increase	Increase the effectiveness of performing required/ desired tasks	Participants want the results to provide insights on how it would increase the effectiveness of end users in performing their required or desired tasks.
	Quicker to complete	Enable me to complete my tasks more quickly	Participants mention the need for the results to demonstrate whether users can complete their tasks more quickly.
	User-control	Give me greater control over my required/ desired tasks	Participants mention the need for the results that enable users to have greater control over their required or desired tasks.
Trustworthy (Kelton et al., 2008)	Accuracy	The extent to which information is free from error	Participants mention the importance of the information in CPM papers being error-free and supported by appropriate data.
	Objective	The degree to which the information is free from bias, deception, and distortion	Participants advocate for the results to be free from bias, deception, and distortion, ensuring that the results present a balanced view.

Theme	Sub-theme	Original definition	Working definition
	Validity	The use of responsible and accepted practice	Participants discuss the importance of using responsible and accepted practices in reporting the results of CPM studies.
	Stability	The information delivered is persistent, both in its presence and its contents, throughout the study results	Participants want the information delivered in the results to be persistent and consistent, in both its presence and contents throughout the study results.

The qualitative analysis focused on identifying the needs of participants regarding understandable, useful, and trustworthy results in CPM studies. CPM results primarily pertain to the Results, Discussion, and Conclusion sections within a CPM manuscript. However, topics that may traditionally belong to the Methods section are also included, provided they are directly tied to the results of the CPM studies. My analysis excluded the content of the interviews referring to other parts of a manuscript, such as writing effective research questions, abstracts, introductions, and methods.

My qualitative analysis began by establishing Inter-Coder Reliability (ICR) between 2 coders (O'Connor and Joffe, 2020). Coder 1 (IR) began the analysis with one transcript by reading and dividing the text into distinct paragraph chunks as the unit of analysis. Each paragraph chunk represented a coding unit. Upon completion of 'chunking' the transcript, the transcript was transferred to a spreadsheet. In this spreadsheet, each row corresponds to an individual paragraph chunk. The columns are organized as follows: 'ID' for identifying each paragraph chunk, 'Text' for the actual content of the paragraph chunk, 'Theme' for the main theme derived from the codebook (i.e., 'understandable,' 'useful' and 'trustworthy') that applies to the chunk, and 'Sub-theme' for the more specific sub-theme within the main theme that the chunk addresses.

During the coding process, Coder 1 assigned codes from the codebook to each paragraph chunk. One paragraph chunk can only have one code. Once this coding phase was completed, coder 1 created a new copy of the spreadsheet containing only the paragraph chunks, without the codes from coder 1, for coder 2 to code the text chunk.

Coder 2 (SM) subsequently applied the codebook to the new spreadsheet. Coder 2 independently coded excerpts of the paragraph chunks using the codebook. The coding process by Coder 2 occurred in several stages. Initially, Coder 2 completed coding for the first two transcripts from two participants and then paused to discuss any discrepancies with Coder 1. This iterative process continued, where Coder 2 finished one more transcript and matched the coding with Coder 1 until a Cohen's Kappa score of at least 0.8 was achieved, indicating a high level of inter-coder reliability or agreement between Coder 1 and Coder 2. We achieved the score after the coding of the 8th interview. At that point, Coder 2 coded the remainder of the data. Both coders still met to discuss when they needed to clarify certain coding parts or the transcript chunk. After completing the coding, Coder 1 used R to assess the final Cohen's Kappa on all transcripts for ICR.

Once Coder 2 completed the coding process and achieved a Cohen’s kappa score of at least 0.8 (O’Connor and Joffe, 2020), Coder 1 reported the results of the analysis that identify needs, which represent what participants want to be included in the reporting of CPM studies to ensure that the results are understandable, useful and trustworthy. One sub-theme can have one or more needs. For example, for knowledge, the sub-theme of understandable has one need. In contrast, the sub-theme ‘analysis’ of the theme ‘understandable’ has five needs. Additionally, Coder 1 separated the paragraph chunks that mentioned topics of visualization preferences, visualization tasks, or visual encodings for further analysis in Chapter 6.

Ethics approval

Participants were informed about the study through consent information in the recruitment email. Furthermore, this study obtained their informed consent for the interviews at the beginning of each interview session. The study was granted exempt status by the Institutional Review Board (IRB) at the University of Washington, Seattle. Upon completion of the interviews, participants received a $25 Tango gift card as a token of appreciation for their time and contributions.

5.4 RESULTS

Participant characteristics

A total of 18 participants completed interviews (ID1 - ID18). Table 5.2 provides an overview of their demographic characteristics, professional experiences, and region based on the descriptive analysis of demographic and experience data reported in the survey from Chapter 4. Most participants fall within the age range of 45 – 65 years old. Like the survey, most were men affiliated with academia in Europe and North America.

Table 5.2 Participant characteristics

Characteristics	n (%) from the interviews	n (%) from the survey
Age	N = 18	N = 263
25 – 34 years old	6 (33%)	46 (17%)
35 – 44 years old	2 (11%)	81 (31%)
45 – 65 years old	10 (56%)	82 (31%)
Prefer not to say	0	2 (1%)
Declined to state	0	43 (16%)
Gender	N = 18	N = 263
Man	12 (67%)	146 (56%)
Woman	5 (28%)	67 (25%)
Prefer not to say	1 (5%)	5 (2%)
Declined to state	0	43 (16%)

Characteristics	n (%) from the interviews	n (%) from the survey
Organization	N = 18	N = 263
Academia	12 (67%)	180 (68%)
Academia & other	3 (17%)	16 (6%)
Other (Government, Consultancy, Freelancing, Industry)	2 (11%)	16 (6%)
Declined to state	1 (5%)	44 (17%)
Region	N = 18	N = 263
Africa	0	7 (3%)
Asia	1 (5%)	27 (10%)
Australasia	2 (11%)	2 (1%)
Europe	5 (28%)	90 (34%)
North America	8 (44%)	61 (23%)
South America	1 (5%)	10 (4%)
Declined to state	1 (5%)	66 (25%)

Table 5.3 describes the participant experiences as authors and reviewers who participated in the interviews. It highlights that most participants identified as both an author and reviewer (n=12), possessing more than 6 years of experience in either role and engaged in both developing and validating models. This table also indicates that no participants identified solely as reviewers.

Table 5.3 Authorship and peer review experiences of participants

Experiences	Identified as an author n (%)	Identified as an author and a reviewer n (%)
Number of studies published or reviewed	N = 18	N = 12
1 – 5 publications	9 (50%)	3 (25%)
6 – 10 publications	2 (11%)	3 (25%)
11 – 20 publications	5 (28%)	3 (25%)
More than 20 publications	2 (11%)	3 (25%)

Experiences	Identified as an author n (%)	Identified as an author and a reviewer n (%)
Years of experience	N = 18	N = 12
1 to 5 years	8 (44%)	2 (16%)
6 to 10 years	3 (17%)	4 (33%)
11 to 20 years	4 (22%)	5 (42%)
More than 20 years	3 (17%)	1 (8%)
Study Type	N = 18	N = 12
Developing and validating models	12 (68%)	10 (84%)
Developing models	4 (22%)	1 (8%)
Validating models	2 (11%)	1 (8%)

5.4.1 *RQ: What needs do biomedical researchers express for CPM study results to be understandable, useful, and trustworthy?*

I conducted the interviews over a period spanning from mid-November 2022 to February 2023. The analysis identified needs, which were grouped into sub-themes of the three a priori themes of ‘understandable,’ ‘useful,’ and ‘trustworthy.’ In total, 37 needs were identified, distributed as 13 needs for 'understandable' across six sub-themes, 11 needs for 'useful' across eight sub-themes, and 13 needs for 'trustworthy' across four sub-themes. Participants' responses showed that 100% (18/18) addressed 'understandable,' 94.4% (17/18) addressed 'useful,' and 100% (18/18) addressed 'trustworthy.' Exemplary quotes for each code (i.e., sub-theme) are provided in the description of each need following the tables (i.e., Table 5.4 – 5.6) that summarize the needs of each theme. Appendix G and Appendix I include the complete transcript text used as the unit of analysis in the quality analysis in this Chapter, along with its corresponding ID, theme, and sub-themes.

5.4.1.1 Understandable study results

Participants identified needs across all six sub-themes within the 'understandable' theme, totaling 13 needs. Each sub-theme encompassed one to four unique needs, with 'analysis' featuring the highest number of needs. Table 5.4 presents a description of participant needs for each sub-theme under the 'understandable' theme, along with their corresponding #Needs. I added a column labelled '#Needs' as an identifier for each need within every sub-theme. The format for #Needs, x-y-z, uses 'x' to indicate the theme (i.e., understandable (1), useful (2), or trustworthy (3)), 'y' for the sub-theme of each theme in sequence, and 'z' for the specific need within each sub-theme, also in sequence.

Table 5.4 Description of needs and the number of needs (#) for each subtheme of 'understandable' study results

Sub-theme	Description of Needs	# Needs
Knowledge	Providing basic information about study results uniformly, adhering to standard practices and guidelines, and ensuring their consistency and transparency. (IDs 1, 2, 3, 12, 19)	1-1-1
Comprehension	Ensuring user-centric reporting and calling for the study results to be understandable to a broad audience, including those with limited technical expertise, especially clinicians as the target users. (IDs 2, 6, 7, 9, 10, 12, 13, 17, 18, 19)	1-2-1
	Making the analysis using machine learning models interpretable that avoid the black box problem, and provide clear, comprehensive explanations of complex methods. (IDs 4, 6, 9, 12, 14, 17)	1-2-2
Analysis	Breaking down the data exploratory phase that delves into the data discovery phase, showcasing initial data signals, and detailing specific data characteristics that could influence model building and interpretation. (IDs 1, 7, 8, 14, 17)	1-3-1
	Describing the detailed steps of model development. (IDs 2, 6, 7, 9, 12, 13, 14, 18)	1-3-2
	Analyzing model performance using multiple metrics. (IDs 1, 3, 6, 7, 11, 13, 14, 18 19, 20)	1-3-3
	Showcasing the complex model relationships between variables and how they contribute to the overall model output and performances. (IDs 2, 3, 6, 8, 10, 13, 14, 19)	1-3-4
Application	Allowing users to try out prediction models through clear operational guidance, user-friendly scoring systems, and the use	1-4-1

Sub-theme	Description of Needs	# Needs
	of web-based applications for active testing and implementation in clinical settings. (IDs 5, 6, 7, 8, 9, 10, 14, 18)	
	Ensuring the applicability of the models in clinical settings, informing treatment decisions, using well-defined and measurable predictors, and being adaptable across different populations in clinical practices. (IDs 5, 7, 10, 11, 14, 17, 18, 19, 20)	1-4-2
Synthesis	Synthesizing the study's contribution to diverse settings in broader research contexts and clinical practices. (IDs 6, 7, 8, 13, 14)	1-5-1
Evaluation	Allowing end users to verify the model results by sharing original data, codes, and analytical processes (reproducible). (IDs 2, 3, 6, 7, 17)	1-6-1
	Acknowledging tradeoffs between models, particularly between robustness and understandability. (IDs 6, 10)	1-6-2
	Comparing the performance of proposed models with the other existing models or standard practices (IDs 6, 19, 20).	1-6-3

Knowledge

#1-1-1 Providing basic information about study results uniformly

Participants described the importance of presenting results of CPM studies with consistent uniformity and in alignment with established standards (IDs 1, 3, 2). This uniform presentation, repeatedly spotlighted as a fundamental aspect, assures clarity and facilitates understanding, particularly for those engaging in systematic reviews or other scholarly pursuits (IDs 3, 12). By adhering to these standardized practices, the information meets audience needs and fosters a shared understanding within the research community, making the results readily comprehensible across various audiences (IDs 19, 2). Participants indicated that such guidelines serve as benchmarks for study reporting and alignment with well-defined criteria of quality (IDs 2, 12, 3), potentially amplifying the study's acceptance within biomedical researcher communities. Thus, uniformity and adherence to standards not only ensure that results meet shared needs but also enhance the overall transparency of the research. As one participant summarized:

"*And then, so the understanding is basically related to the standardized or uniformity of the information that is already there. You should report that, and in a way that is already acceptable in the community of the research community*." (ID 3)

Comprehension

#1-2-1 Ensuring user-centric reporting

Participants underscored the importance of user-centric reporting of the study results of CPM studies, emphasizing the necessity for predictions that are not only scientifically rigorous but also tailored to a diverse audience (IDs 6, 7, 12, 17, 18, 2, 10). This approach includes crafting results comprehensible to clinicians, healthcare policymakers, and other non-specialist readers. Participants acknowledged that failure to make these complex models accessible could lead to them being disregarded (IDs 12, 9). As one participant summarized:

"*So, clinicians sometimes will not use these prediction models even though they may be backed by data... because they don't understand the model itself. They don't know how it works."* (ID 9)

Crafting the papers without challenges necessitates collaboration between statisticians and clinicians, a multi-disciplinary approach emphasized by participants (IDs 13, 19). Their collaboration ensures a comprehensive understanding of the methodology and its clinical significance, bridging the gap between technicality and practicality. Such a multi-disciplinary approach offers a richer, more balanced portrayal of the study's findings, catering to a diverse readership. This collaboration was evident in the statement:

"*And usually, they (clinicians and statisticians) talk different languages... And therefore it, so you need both competences to have a good prediction model. Even in planning development of such a model.*" (ID 13)

#1-2-2 Making the analysis using machine learning models interpretable

Participants consistently emphasized the necessity of enhancing the interpretability of machine learning analysis used in CPM studies (IDs 6, 9, 12 14, 17). They expressed a shared conviction that comprehension of the complex analysis must remain accessible to a broad audience (IDs 4, 12). A recurring concern centered on the tendency to treat machine learning or deep learning models as "black boxes" that are difficult to interpret (IDs 17, 6). This sentiment laid a foundation for a broader discussion on the challenges of analysis and the imperative of explicit communication. As one participant noted, emphasizing the urgent need for more intelligible explanations:

> " *I think what would help understandability is a clear statement of the point in the progress of their study, where, you know, that has gone into, you know, the black box and what the method is that has been used. And some easily understood brief explanation of what that might entail, you know, whether it is, you know, this is clearly a method where the computer has been taught to recognize certain patterns, or whether this is a machine-generated model on the basis of the information that's there, just something that's a bit more explicit, rather than what a lot of the papers just assume that, well, we've taken these variables, we fed them in, and out they come at the other end. So a bit more explicitness, I think.*" (ID 12)

Analysis

#1-3-1 Breaking down data exploratory phase

Participants emphasized presenting the results from an exhaustive data exploration during the discovery phases (IDs 8, 17). They valued the transparency of showcasing initial data signals and highlighted their challenges in deciphering variable patterns (IDs 14, 7). Furthermore, the participants noted the necessity of detailing specific data characteristics (IDs 7, 17). They stressed

the importance of understanding how certain properties and characteristics could impact the building and interpretation of predictive models (IDs 14, 1). For example, one participant pointed out that knowing the distribution of a variable could help in selecting appropriate transformation techniques and avoid misleading interpretations. This illustrates the critical role of initial data exploration in setting the stage for robust and reliable predictive modeling, ensuring that the foundational assumptions and data properties are well-understood: A participant captured this sentiment, stating:

> " *I think this is very important, especially in the data discovery phase. First of all, to show the readers what signals the data may have in the beginning, just to show what kind of characteristics the data has, how much variability there is there in each of the features that's being questioned.* " (ID 17)

#1-3-2 Describing the detailed steps of the model development

Participants consistently voiced the need for results of CPM studies to give a comprehensive description of the analysis during model development. The analysis reporting should dissect the model-development process into distinguishable steps while simultaneously detailing the connections between each step that lead to the final prediction models (IDs 6, 12, 17, 18). This description should encompass the selection of predictions, modeling process (e.g., logistic regressions, random forest), and the refinement of the models that may be attempted to achieve model performance (IDs 2, 7, 9, 13, 14, 18). This needs highlights for full analysis reporting during model development, as eloquently expressed by one participant:

> "*It seems like just good research, you know, components of good research that, you know, your goals are very well stated, your, your variables are, are very well defined, that your methods are well defined. Whatever, you know, how did you, how did you select these*

variables? What were your, your, you know, what was your variable selection procedures, your validation procedures, all that kind of stuff. So, all that stuff that needs to be very well defined." (ID 14)

#1-3-3 Analyzing model performance using multiple metrics

Participants also advocated using diverse, robust performance metrics and tools to assess model performance. Measurements such as ROC, Breyer scores, and calibration curves were suggested for the complete depiction of model performance (IDs 1, 3, 6, 8, 19, 7, 13, 14, 11, 20). It was this holistic view that would allow for a better understanding of the model performance. Participants emphasized that relying on a single metric is insufficient to capture the full picture of model performance. For instance, while the ROC curve provides insight into the model's discriminatory ability, it does not address calibration, which is equally important. Calibration curves help to ensure that the predicted probabilities accurately reflect the actual outcomes. Additionally, Breyer scores can provide a different perspective by quantifying the overall performance of probabilistic predictions. This highlights the necessity of multiple performance metrics to provide a comprehensive evaluation, ensuring both the accuracy and reliability of the predictions made by the model. One participant captured this sentiment succinctly when they said,

" *The other aspect of clinical prediction models is that there must always be a calibration curves. And so, I insist on them because otherwise people think they can get a probability and it is meaningful. And that may have a very good discrimination but poor calibration*." (ID 19)

#1-3-4 Showcasing the complex model relationships between variables

Another perspective centered on the need for presenting model details, particularly the complex relationships between variables and how they contribute to the overall model results, is

vital (IDs 3, 8, 10, 13, 14, 19). The relationships should highlight how model performance changes under the conditions of covariate variations, which could provide a complete understanding of the models' function, possible limitations, and models' robustness (IDs 2, 6, 14). A covariate is an independent variable that can influence the outcome of the model and is typically controlled or measured to assess its impact. For example, understanding how different levels of a covariate such as age impact the model can reveal how the model performs across different subgroups of the population. This can be particularly important in ensuring that the model is equitable and performs well not just on average, but for all relevant subgroups. As one participant insightfully summarized,

> "*If you hold everything fixed at a certain value and then vary age, you know, because 18-year-olds probably have different, you know, number of comorbidities or something like that, than 90-year-olds. And so, by holding everything constant, you know, it is still useful to see what the shape looks like ... you know, (model) performance.*" (ID 6)

Application

#1-4-1 Allowing users to try out prediction models

Participants articulated a strong inclination towards experiencing prediction models in action, emphasizing the need not just to read about them but to actively engage with them (IDs 5, 6, 9, 14, 8, 7,18). This desire bridges the gap between theoretical articulations and hands-on usability, with an emphasis on clear operational guidance, transparent scoring systems, and user-friendly tools that ensure models are both accurate and comprehensible (IDs 6, 8). For participants, understanding the inner workings of a CPM is essential for its successful implementation in clinical settings (IDs 6, 9, 14). The digital era further presents opportunities for hands-on engagement, with a trend towards leveraging online platforms, especially web-based applications,

to facilitate interactive experiences and ensure users transition from passive consumers to active contributors (IDs 7,10, 14, 18). As one participant described:

> "*Actually, figures to make it more communicable to readers. For example, for studies of COVID-19, how to calculate the model and apply the results. Those figures are necessary?"* (ID 5)

#1-4-2 Ensuring the applicability of the models in clinical settings

Participants emphasized the importance of model applicability in clinical settings, underscoring that utility hinges on both robustness and integration into clinical routines (IDs 7, 11, 14, 5, 10, 19, 20). They argued for optimal use in clinical contexts (ID 19), clear definition and measurability of predictors (IDs 11, 14), and adaptability across different populations in clinical settings to ensure wide-reaching relevance (IDs 17, 18). The discussion also highlighted the need for models to incorporate easily measurable variables and model flexibilities, enabling a deeper grasp of model application in their clinical practices (IDs 5, 10, 17). For example, a model designed to predict cardiovascular risk must not only provide accurate predictions but also seamlessly fit into existing clinical workflows, such as integrating with electronic health records (EHRs) to automatically flag high-risk patients for further assessment. Additionally, predictors like blood pressure or cholesterol levels should be readily measurable in a typical clinical visit to facilitate swift decision-making. Another consideration is the model's performance across diverse patient populations, ensuring that it remains accurate for different age groups, ethnicities, and comorbidities, thus making it a versatile tool in varied clinical scenarios. As one participant aptly summarized:

> "*And some discussion of how the model would be used in clinical care. You know, what part of the calibration curve are you actually looking at? Do you want to treat people who*

are very high risk? Are you trying to exclude people who are low risk? You know, what exactly is it that you're planning to use this model for? And then just saying, we have a model to identify risk factor. You know, we've identified risk factors for a particular disease." (ID 20)

Synthesis

#1-5-1 Synthesizing the study contribution to diverse settings

Participants echoed a consistent desire to report the results of CPM studies to proactively and strategically narrate how the research provides innovative contributions to the broader landscape (IDs 8, 13, 6, 7, 8, 14). The essence of impactful research lies in the authors' ability to provide a clear narrative about how their findings contribute to the broader field, contextualizing the study and offering readers a glimpse into its overarching significance (IDs 8, 13). Additionally, participants emphasized the need for patient outcomes to be at the forefront of clinical research, aligning findings with patient-centered outcomes and fostering an understanding of the patient journey. The emphasis on a bidirectional flow of information ensures that research is theoretical and rooted in real-world clinical implications, focusing on tangible impacts on care quality (IDs 6, 7, 8, 14). As one participant insightfully summarized:

"*The authors should draw a clear line, tell a clear narrative of what this result, what line of research this result is contributing to, and what the overall goal of that type of research might be.*" (ID 8)

Evaluation

#1-6-1 Allowing end users to verify the model results

Participants stressed the importance of enabling those who read the result report of CPM studies to double-check the prediction models' results highlighted in the studies (IDs 3, 7). The

keyword here was "reproducibility," where authors are expected to provide access to the original dataset, analysis flows, and codes as a standard practice (IDs 3, 6, 7, 2, 17). This approach ensures other experts can test and confirm the results for themselves. For instance, if a CPM study claims that a particular model can predict patient outcomes with high accuracy, providing the raw data and analysis scripts enables other researchers to verify these claims independently. Moreover, reproducibility helps identify potential errors or biases in the original analysis, promoting continuous improvement in model development and application. This approach means that other researchers can test and confirm the results for themselves. One participant mentioned this:

> *"To suggest that authors need to provide both the raw data set and the analysis flow alongside the manuscript...It is claimed that other people can then access and use it. So maybe attaching that would be some kind of almost an indirect element of quality, but it kind of pushes you to be more careful, to be more structured."* (ID 2)

#1-6-2 Acknowledging tradeoffs between models

Participants underscored the essential need for the results of CPM studies to be clear about the model's tradeoffs, particularly between robustness and understandability (IDs 6, 10). This clarity facilitates a more accurate assessment of the real value of the proposed CPMs. Participants emphasized the importance of acknowledging that CPM study reports sometimes present models in a way that might risk the robustness or accuracy of the models. For example, in CPMs that use logistic regression, converting coefficients to scores would make the model more understandable but reduce its predictive accuracy. This tradeoff is dilemmatic because a highly understandable model that lacks robustness might lead to suboptimal clinical decisions, while a robust but complex model might be too difficult for practitioners to use effectively. Thus, authors must weigh these

tradeoffs to ensure that the chosen model aligns with the intended clinical application and user capabilities. One participant succinctly expressed this sentiment:

> "*I don't think simple models are good. Sort of, but they're also. Sometimes, there's all kinds of tradeoffs both potentially accuracy versus understandability but even if something's understandable*." (ID 6)

#1-6-3 Comparing the performance of proposed models

Participants expressed the need to judge the performances of the proposed prediction models by contrasting them directly against each other and with standard clinical practices (IDs 20, 19, 6). This approach provides a clear perspective on the models' relative value and effectiveness, enabling an evidence-based assessment of whether the proposed models offer advantages over existing diagnostic or prognostic methods. For example, comparing a new CPM with traditional methods such as clinician judgment or existing CPM can reveal whether the new model improves prediction accuracy, patient outcomes, or decision-making efficiency. Such comparisons offer a tangible and practical lens through which the value and application of a model can be ascertained. Without these direct comparisons, clinicians may find it challenging to adopt new models over tried-and-tested methods. Additionally, this comparative approach can highlight specific areas where a model excels or falls short, guiding further refinement and optimization. As one participant noted:

> *"They're not going to tell you how that's going to influence your care. You know, what are the outcomes of your patients? How are they going to be changed by using this model or not using that model? Those pieces of information are never in those papers… The papers could be written in a way that the physicians would understand other than to compare use of the model to clinical practice*." (ID 20)

5.4.1.2 Useful study results

Participants reported 11 needs for all 8 sub-themes within the theme ‘useful.' Table 5.5 summarizes the needs identified by participants within the 'understandable' theme, detailing each need with its associated #Need identifier.

Table 5.5 Description of needs and the number of needs (#) for each subtheme of ‘useful’ study results

Sub-theme	Description of Needs	# Needs
Critical support	Guiding whether or not to use the prediction models by appropriately clarifying their intended function and outlining potential limitations and risks to avoid misuse. (IDs 2, 3, 6, 7, 10, 11, 13, 17)	2-1-1
	Cautioning the unintended consequences of misappropriately using the prediction models, including potential pitfalls, ethical concerns, healthcare burdens, and potential treatment deterrence due to the model predictions. (IDs 1, 3, 6, 10, 11)	2-1-2
Easy to use	Formatting CPM for ease of use among users by portraying the models as straightforward and user-friendly, typically through intuitive tools like nomograms and interactive digital platforms. (IDs 6, 10, 11, 13, 18, 19, 20)	2-2-1
More accomplishment	Informing future research that involves identifying new primary risk factors and suggesting new populations or areas of interest for follow-up research. (IDs 6, 14, 18, 19)	2-3-1
	Offering specific implementation strategies for practical integration of prediction models into clinical practice. (IDs 9, 10, 11, 13, 19, 20)	2-3-2
Quality increase	Suggesting that the prediction models should demonstrate improvements in patient outcomes through rigorous impact analysis studies such as randomized controlled trials (RCTs). (IDs 2, 6, 7, 14, 20)	2-4-1
Useful for tasks	Demonstrating the models’ robustness and how the models work. (IDs 3, 6, 10, 13, 17, 18)	2-5-1
	Ensuring users that the models have practical applicability and utility in real-world clinical scenarios. (IDs 6, 11, 19).	2-5-2
Performance increase	Demonstrating how the prediction models would enhance clinicians' performances by facilitating better-informed decisions and seamlessly integrating into their clinical workflows. (IDs 6, 7, 10, 11, 20)	2-6-1
Quicker to complete	Enhancing content for quick readability by structuring the study results for rapid comprehension and extraction of core insights, with a balance between concise text summaries and complementary visuals. (IDs 8, 19)	2-7-1

Sub-theme	Description of Needs	# Needs
User-control	Empowering readers to have more control over their required or desired tasks, suggesting the incorporation of interactive simulations and hypothetical scenario analyses for enhanced engagement with the study findings. (IDs 2, 6, 7).	2-8-1

Critical support

#2-1-1 Guiding whether to use or not use the model

Participants consistently emphasized the need to define when and how CPMs should be utilized, pointing to both their pivotal role and the potential dangers of misuse (IDs 17, 11, 2, 3, 13, 6, 7). They stressed the tangible advantages of models in high-stakes scenarios, such as determining surgical immediacy or assessing prolonged medical treatments based on risk factors (IDs 17, 11). In another issue, participants firmly advised against misusing prediction models as causal interpretation tools (IDs 6, 7, 13). Participants also highlighted the limitations of risk prediction models that focus on non-modifiable risk factors, such as past medical events, emphasizing that these models primarily inform about risk but do not necessarily suggest actionable interventions for risk mitigation (ID 10). Furthermore, they also called for papers to detail reservations about the models, such as concerns about adaptability during contextual shifts, transparency in "black box" systems, and instances where assessments beyond certain thresholds were deemed inadequate (IDs 2, 3, 13). One participant's statement encapsulates this caution, highlighting the need for contextual awareness:

> "*This is all based on the data that we have, and it will only work under the assumption of no major contextual change in the system because if the system changes, then the entire model kind of goes away. So I guess that the post hoc explanation of the model is something that should be developed more.* " (ID 2)

#2-1-2 Cautioning the unintended consequences

Participants emphasized the vital need to address the unintended consequences and broader implications of CPM adoption (IDs 3, 6, 10, 1, 11). The insights underscored the importance of clarity in research papers, warning against the dangers tied to model misinterpretation and the ethical considerations in their application (IDs 3, 6, 11). For example, a model that inaccurately predicts patient outcomes could lead to over-treatment or under-treatment, potentially causing harm to patients. Ethical considerations include ensuring that the model does not inadvertently reinforce existing biases or inequalities in healthcare. Participants also stressed the need for a balanced approach in applying predictive models, weighing risk-benefit considerations in treatments, and recognizing the potential societal impacts (IDs 1, 11). This includes evaluating whether the benefits of using a predictive model outweigh the risks, and considering how the model might affect different patient populations. These concerns culminated in a participant's observation, which succinctly encapsulates the core theme:

> "*And if you not treat anybody, (or) if you treat anybody, all the person you treat with some thresholds* (e.g., risk score cutoffs, probability thresholds). *So, what they did is nothing was like in this in limits. It was actually going beyond the limits. It means that the model is actually harmful. And but the author hasn't really interpreted that well to communicate that this model is actually useless. It is just rubbish*." (ID 3)

Easy to use

#2-2-1 Formatting CPM for ease of use among users

Participants described the importance of presenting the results of CPM studies to emphasize their ease of use and intuitive application, making them accessible to end-users (IDs 6, 18, 11, 20, 10, 13, 19). The discussion gravitated towards formats that facilitate actionable

outcomes (ID 18), such as the use of nomograms and point-based scores (ID 11) or direct risk estimation that transcends traditional point systems (ID 20). The move towards digital platforms like web applications was highlighted, with these tools seen as particularly effective for enhancing user-friendly interaction and transparency (IDs 10, 13). Platforms such as MedCalc and Shiny apps were appreciated for their ability to allow users to enter data and receive immediate probabilistic outcomes or graphical representations (IDs 18, 19). Reflecting on this sentiment, one participant aptly noted:

> "*Yeah. I don't know if I will use the word appealing, but I will use the word useful. Useful. Or usable. (inaudible). You can do something else like a risk chart. You can have it print or some tool that actually makes my life easier as an end user*." (ID 18)

More accomplishment

#2-3-1 Informing future research

Participants consistently emphasized the multifaceted role of reporting results of CPM studies in not only presenting findings but also initiating subsequent research and delineating future investigative pursuits and follow-up (IDs 18, 19). This holistic view regards an ideal paper as one that builds upon the present to create a roadmap for the future, shaping research trajectories through the identification and in-depth exploration of primary risk factors that can be carried forward into future research (IDs 6). One participant encapsulated this idea by stating:

> "*I think of one as just sort of identifying the largest risk factors in terms of their effect magnitude... if you recognize the risk factors, they can suggest future research directions*." (ID 6)

Further, CPM papers were seen as catalysts energizing subsequent research efforts, with well-articulated results capable of stimulating the scientific community to probe overlooked nuances (IDs 19, 14). As one participant aptly concluded:

> "*You might create a clinical prediction model, which then sparks further work, bringing to light overlooked patient characteristics. Such a model, in turn, can potentially stimulate more research*." (ID 14)

#2-3-2 Offering implementation strategies for practical integration

Participants consistently emphasized the necessity for reporting results of CPM studies to move beyond mere data presentation, focusing on the real-world needs and practicalities of clinicians (IDs 20, 19, 11). They advocated for models that offer tangible pathways to integrate findings into clinical practices (ID 13). This practical approach should stimulate further examination within practitioners' datasets, potentially refining existing care pathways (ID 19). Participants also highlighted the importance of authors demonstrating a comprehensive understanding of the clinical landscape to ensure the model's adoption (ID 9). As one participant aptly stated:

> "*All I can say is that I would just put that as a if authors want people to use their model, they need to understand the conditions under which this will be adopted*." (ID 9)

Quality increase

#2-4-1 Suggesting that the prediction models should demonstrate patient outcomes

Participants emphasized the need for reporting results of CPM studies to demonstrate real-world applicability and tangible influence on improving patient outcomes (IDs 7, 14, 2, 20). For instance, a CPM that predicts the risk of complications after surgery should provide evidence that its use leads to fewer complications and better recovery rates. They also stressed the importance of

assurance regarding a model's demonstrations to improve the outcomes by encouraging the use of randomized controlled trials (RCTs) for validating the model outcomes (ID 6). RCTs provide robust evidence by comparing patient outcomes with and without the use of the model, thereby confirming its effectiveness. Ultimately, the consensus was that an effective model must show evidence of its influence on patient care, aptly summarized by one participant:

> *"I'm not sure that usefulness is completely achieved because it doesn't say just one small piece of the total puzzle. It doesn't say whether the use of that prediction model would actually lead to improving decision making and thereby the patient's outcome."* (ID 7)

Useful for tasks

#2-5-1 Demonstrating the models' robustness and how the models work

For participants, the primary focus of CPM study results should be to support readers by demonstrating the models' robustness and how the models work. To them, reporting CPM study results should effectively communicate comprehensive model performance, including how the model performances vary at specific levels (IDs 7, 18), and model robustness in multiple scenarios (IDs 3, 10). For example, a model predicting the risk of diabetes should show consistent performance across different age groups and ethnicities, ensuring it is reliable under various conditions. Furthermore, regarding how the models work, participants expect the study results to reveal influential predictors or variables in the model and demonstrate how different predictors or their combinations influence the results (IDs 6, 13, 17). This could include detailing how factors like BMI, diet, and genetic predisposition contribute to the model's predictions, providing transparency and insight into the model's functionality. Collectively, this need underscores how CPMs can better serve readers by providing a clear understanding of the model's strengths,

limitations, and practical applications. Collectively, this need underscores how CPMs can better serve readers. One participant succinctly encapsulated this sentiment by stating:

> "*I know the third part of your talk will be on robustness, statistical robustness, but I think that usefulness is also part; it is intrinsically part of how robust the model is. (inaudible) What's the quality of evidence? So, I guess that will, it is intimately linked*." (ID 10)

#2-5-2 Ensuring users that the models have practical applicability

Participants emphasized the imperative of reporting CPM study results to demonstrate theoretical effectiveness and ensure practical applicability and utility in real-world clinical scenarios (ID 11). This highlighted that demonstrating practical applicability means showing that a model works not just in controlled research settings but also in the varied and unpredictable environments of daily clinical practice. The necessity for thorough validation surfaced prominently, with participants noting that too many papers focus on model derivation without the crucial follow-up validation and studies across multiple centers, thereby casting doubts on their real-world generalizability (IDs 11, 19). For example, a model developed in one hospital should be tested in various other hospitals to confirm its broader applicability. The flexibility of the model to adapt to diverse and complex clinical scenarios in day-to-day practices emerged as a concern, intertwined with inquiries about the model's specific or generalized application in clinical settings (IDs 6). This includes evaluating whether the model can handle different types of patients and varying clinical workflows without significant loss of performance. One participant shared this apprehension:

> "*And is there any flexibility in the model? How much flexible is this model that can accommodate the complex (clinical) scenarios? So, these kinds of things are still lacking in the literature*." (ID 11)

Performance increase

#2-6-1 Demonstrating how the prediction models would enhance clinicians' performance

Participants described the important role of CPM papers in enhancing clinicians' performance beyond mere statistical presentations, expressing a need for actionable tools to aid in improved patient care and better-informed decisions that could be integrated into everyday clinical routines (IDs 20, 7, 10, 11, 6). The emphasis was on models that bolster decision-making processes (IDs 20, 7), with many highlighting the value of clear, actionable insights like decision trees or algorithms (ID 20). For instance, a decision tree could help a clinician quickly determine the best course of action based on patient-specific variables. Another key discussion point was the importance of integrating these models into regular clinical workflows to ensure their real-world applicability and alignment with clinical realities (IDs 6, 7, 10). This means the models should be designed to fit seamlessly into existing systems and processes, such as EHR, to avoid disrupting the clinicians' workflow. The overarching theme, encapsulated by one participant, demonstrated how the prediction models could facilitate clinicians' work:

> "*I would say useful results are results that can inform the usability of the prediction model. And in that sense, I would say it should inform on the degree to which the decision-making in clinical practice would be affected*." (ID 7)

Quicker to complete tasks

#2-7-1 Enhancing content for quick readability

Participants expressed a desire for reporting results of CPM studies to be structured in a manner that facilitates rapid comprehension and extraction of core insights (IDs 8, 19). Striking a balance between concise text summaries and complementary visuals can further streamline the reading process; text should prioritize essential insights while graphs delve deeper, satisfying both

readers in a hurry and those seeking depth (IDs 19, 8). Ultimately, the aim, as voiced by participants, is for readers to efficiently distill the paper's main conclusions into a few sentences, epitomized by one statement:

> "*I should be able to walk away from the paper and summarize that useful result in just a couple of sentences."* (ID 8)

User-control

#2-8-1 Offering greater control and interactivity with study results

Participants expressed a strong preference for CPM results that enhance user interaction and provide in-depth control over the analysis. They highlighted the utility of interactive simulations as a key method for achieving a more comprehensive understanding of the data. By allowing the visualization of how individual predictors influence performance metrics, such as AUC or MSC, users can discern the importance of each variable (ID 6). Furthermore, the discussion extended to the value of creating hypothetical scenarios in which changing model parameters or underlying assumptions lead to different predicted outcomes. This feature of dynamic simulations, as emphasized by the participants, not only illuminates critical factors but also engages readers by demonstrating the practical applications of theoretical models (IDs 2, 7). Through this interactive approach, the results empower users to have more control over their required or desired tasks as readers, thereby enhancing engagement with the study findings and linking abstract concepts with tangible, real-world scenarios, as one participant eloquently noted.:

> "*What I would like to see maybe some sort of additional analysis or simulations that could illustrate the use of the prediction models in real life. So now we have the prediction models. If it is used in that way, then the impact on the patient outcome will be that.*" (ID 7)

5.4.1.3 Trustworthy study results

I identified 13 needs from four sub-themes under the theme category of 'trustworthy.' Table 5.6 provides a comprehensive overview of the trustworthy theme, outlining participant needs and their respective #Need identifiers for further reference and analysis.

Table 5.6 Description of needs and the number of needs (#) for each subtheme of 'trustworthy' study results

Sub-theme	Needs	# Needs
Accuracy	Ensuring accuracy in data presentation that avoids mistakes such as discrepancies in displayed results and misleading visual representations. (IDs 3, 8, 11, 12, 17, 19)	3-1-1
Objective	Acknowledging limitations inherent in study results that include acknowledging the importance of confidence intervals over point estimates, addressing potential biases, and underscoring the role of calibration, alongside discrimination. (IDs 8, 12, 17, 19, 20)	3-2-1
	Ensuring that the study results are reproducible by encouraging transparency, open science frameworks, online code sharing, and structured data-sharing mechanisms. (IDs 4, 7, 8, 10 17)	3-2-2
	Justifying the decisions made during the research process, which includes the careful selection of analytical methods and variables, confronting and addressing anomalies like outliers, detailed methodology disclosure, handling missing data, determining sample size, and avoiding overly aggressive methods. (IDs 1, 2, 3, 8, 12, 13, 14, 17, 18)	3-2-3
	Substantiating the model performance and improvement claimed using multi-dimensional assessment methods and benchmarking against existing models. (IDs 2, 6, 10, 19, 20)	3-2-4
Validity	Reporting study results adhered to standard practices. (IDs 1, 3, 9, 10, 11, 18)	3-3-1
	Applying proper validation to the models. (IDs 2, 11, 13, 14, 20)	3-3-2
	Upholding the adherence to the study's method and goals. (IDs 4, 8, 14, 18, 19)	3-3-3
	Demonstrating model replicability across various contexts, indicating an essential marker of the model's reliability and validity (IDs 2, 4, 9, 12)	3-3-4
	Recognizing the limitations of established guidelines and adjusting accordingly to maintain methodological soundness. (IDs 1, 3, 10)	3-3-5

Sub-theme	Needs	# Needs
	Selecting representative samples, ensuring alignment with the broader group of the intended application, and safeguarding against potential bias. (IDs 7, 13, 14, 18)	3-3-6
	Using trustworthy datasets encompassing datasets that are validated and standardized variables. (IDs 8, 10, 11, 12, 18, 20)	3-3-7
Stability	Demonstrating model equity across individuals and populations. (IDs: 7, 12)	3-4-1

Accuracy

#3-1-1 Ensuring accuracy in data presentation

Participants strongly advocated for impeccable accuracy in data presentation within the results of CPM studies to ensure that the study results were trustworthy. Mistakes, like misrepresenting percentages without context or discrepancies in results, can undermine confidence in the findings (IDs 3, 19). Accurate visual data representation, including error bars and graphs, is crucial, as misleading visuals can significantly distort data interpretation (ID 8). Additionally, avoiding inaccurate or defensive statements about the results of the CPM studies should be avoided (IDs 17, 12, 11). As one participant incisively observed:

> "*I didn't trust what they, how they were presenting the results, because they presented a table with the true positives, false negatives, true negatives, false positives, sensitivity, NPV, all these things. And it was quite obviously wrong, the sensitivity, because they would present a sensitivity of 90 something%, but they had no false negatives, which means it must be 100%. So, I see that sort of thing sometimes myself.*" (ID 11)

Objective

#3-2-1 Acknowledging limitations inherent in study results

Participants emphasized the necessity for an unbiased, deception-free, and clear presentation of results in CPM studies by discussing thoroughly the limitations of the results (IDs

8, 12). They highlighted the importance of addressing inherent study limitations, which includes acknowledging potential biases, detailing data cleaning and preprocessing practices, and integrating confidence intervals into the analysis (IDs 19, 17). For example, if a study's sample is not representative of the broader population, this should be clearly stated, along with any steps taken to mitigate this bias. Incorporating limitations, such as favoring confidence intervals over point estimates, is crucial as it enhances trust among end users (ID 19). Confidence intervals provide a range of values within which the true effect likely lies, offering a more nuanced understanding than a single point estimate. By transparently discussing these aspects, authors can convey a more accurate picture of the model's reliability and applicability. Collectively, these insights support the notion that trustworthy research thoroughly addresses limitations, as one participant aptly summarized:

> "*They heavily discuss the limitations of the research and are transparent about what those are. I think that trustworthy research is something that, you know, I shouldn't walk away from reading that paper thinking, wow, they really missed out on this important limitation*." (ID 8)

#3-2-2 Ensuring that the study results are reproducible

Participants highlighted the delicate balance between transparency and the challenges associated with sharing data and codes in reporting results of CPM studies, emphasizing the complexity of achieving reproducibility. They regarded reproducibility as essential for reporting trustworthy results in CPM studies, which involves duplicating results using identified risk factors, the original data, codes, and proposed models (IDs 10, 4, 7). Despite recognizing the difficulties in sharing data and complex model-building codes due to intellectual property, they favored transparency, especially as methodologies evolve (IDs 4, 7). They also supported using open

science frameworks and detailed data preprocessing to enhance reproducibility and reduce biases (IDs 8, 17). The conversation strongly supported sharing the code of the prediction models online, aligning with the principles of scientific transparency and highlighting platforms like GitHub as key to this shift (IDs 4, 7). This collective call for more structured data-sharing practices culminated in a participant's reflection, stating:

> "*I think maybe at the very beginning when I started research, I was more resistant to that. But now I think it is better for everyone to have the codes (of the prediction models) online."* (ID 7)

#3-2-3 Justifying the decisions made

Participants consistently emphasized the need for reporting rigor and justification in all decisions made throughout the research process of CPM studies, including the careful selection of analytical methods and variables to develop the models (IDs 8, 12, 18, 17), addressing anomalies like outliers and models' assumptions (IDs 3, 18, 2, 17), handling missing data (IDs 1, 8, 13), determining sample size (IDs 3, 14, 17), and avoiding overly aggressive methods (ID 2). They believe that justifying these decisions not only maintains the integrity of the research but also fosters reliability among readers (ID 13). A shared sentiment among participants stressed the necessity for transparent decisions in research actions, such as selecting analytical methods or variables, ensuring that each choice was rooted in sound reasoning (IDs 8, 12, 3). Addressing anomalies, including visualized outliers, emerged as a critical concern, with participants underscoring the need to confront and justify these unusual aspects (IDs 2, 18, 3). Justifying the sample size was another central theme, with the need for substantial size to ensure statistical relevance and caution against problems like the "curse of dimensionality" when inadequate (IDs 13, 14, 17). Participants also highlighted the significance of handling missing data with responsible

practices (IDs 8, 13). In addition, concerns arose about the risks of overly aggressive approaches to achieve high performances and the pitfalls of unplanned model selection (ID 2). This intricate theme culminated in a reflection by one participant:

> "*Well, that is a problem because when people do work on a question like this, they want some kind of predictive model, right? They want a significant result. But the problem is that if they don't get what they expected, then they start panically asking for more sophisticated models... they're causing overinflation of false results*." (ID 2)

#3-2-4 Substantiating the model performance and improvement claimed

Participants emphasized the importance of substantiating claims regarding model improvement to ensure integrity and unbiased findings. They highlighted the need for multifaceted evaluation, cautioning against reliance on a single perspective, such as using only metrics like AUC and not including calibration measures as well as confidence intervals, when claiming model performance (IDs 6, 19, 20). For example, while AUC (Area Under the Curve) measures the ability of a model to distinguish between classes, calibration measures how well predicted probabilities agree with actual outcomes, providing a more comprehensive assessment of model performance. On another front, they highlighted the importance of rigorous benchmarking against existing models and testing the models' robustness, supported by transparent reporting and methods (IDs 6, 2). This involves comparing the new model with established models to demonstrate its relative effectiveness and robustness across different datasets and conditions. This reporting ensures that other researchers can replicate the findings and validate the claims made. As one participant remarked:

> "*If you come up with a new one, you should compare it to the one everybody is using and not just present your results on their own.*" (ID 6)

Validity

#3-3-1 Reporting study results adhered to standard practices

Participants strongly emphasized the importance of adhering to recognized practices and standards when reporting CPM study results. Emphasizing complete adherence to established guidelines and protocols contributes to consistent quality and a rigorous scientific methodology in the results of CPM studies (IDs 18, 10). Furthermore, adherence to standardized practices, guided by various guidelines and checklists, ensures more trustworthy results, especially in terms of their validity (IDs 3, 10, 1, 9, 11). Reflecting on this perspective, one participant noted:

> "*So, of course, results are not 100 percent perfect. That's why I said that when I read the paper, I feel like how much efforts are being done and then how they can improve those results by using some like either the TRIPOD or there are some other guidelines in terms of assessing the biasness of the studies*." (ID 3)

#3-3-2 Applying proper validation to the models

Participants emphasized the importance of robust validation in establishing trustworthy results of CPM studies and ensuring their practical application (IDs 2, 11, 20). Some expressed concerns about models that had been in use for extended periods without undergoing proper validation, while others underscored the need for validation to be specifically tailored to the intended population to avoid inaccurate results (IDs 14, 13). For instance, a model developed using data from one population might not perform well when applied to a different demographic group, emphasizing the need for context-specific validation. Echoing calls for stringent validation procedures, one participant noted:

"*I sometimes saw papers when they applied the scoring system and did not consider that the population was not according to the original one (the population in which the prediction model was developed).*" (ID 13)

#3-3-3 Upholding the adherence to the study method and goals

Participants emphasized the critical role of robust and transparent results in CPM studies that adhere to their methods and goals, ensuring the paper's validity. They championed aligning methodology with the study's goals and underscored the necessity of presenting clear details of the study pipeline (IDs 14, 8). Participants' confidence in the paper was rooted in a meticulously detailed and repeatable methodology, eliminating gaps or ambiguities and establishing a foundation for responsible and accepted practices (IDs 19, 4, 18). Highlighting the importance of reporting the study results that uphold the adherence to the methods, as one participant noted:

"*I do think that possibly even having a table where they go step by step in order of the analytic procedures that were done. Like maybe first, the data were checked for normality and skewness. And then second, the data were log-transformed."* (ID 8)

#3-3-4 Demonstrating model replicability

Participants emphasized the importance of demonstrating the replicability of CPM study results as an essential marker of validity (IDs 2, 4). Replicability differs from reproducibility in that it involves replicating results using different contexts and datasets, whereas reproducibility entails using the same data to produce identical results (National Academies of Sciences et al., 2019). They highlighted that without such replication, particularly across independent datasets, the utility and credibility of the model can be seriously questioned (IDs 2, 4, 9, 12). For example, a model that accurately predicts patient outcomes in one hospital should also perform well when tested in different hospitals with varying patient demographics and clinical practices. This cross-context

validation ensures that the model's predictions are not merely a result of specific characteristics of the original dataset. To them, the ability to replicate results in various contexts is an unambiguous testament to a model's reliability and robustness (IDs 9, 12). Reflecting on the crux of this emphasis on replication, one participant observed:

> "*So I guess replication is the key because if the results are replicable, then it means that I can apply them to my situation. If they are not replicable, then you are in a whole field of possibilities what might have gone wrong*." (ID 2)

#3-3-5 Recognizing the limitations of established guidelines and adjusting accordingly

While established standards and guidelines serve as robust roadmaps for researchers, participants also highlighted the importance of being aware of their limitations (IDs 1, 3, 10). Not all guidelines are perfect or exhaustive, and their applicability may vary across studies. It is imperative for researchers to recognize these shortcomings, evaluate their relevance to their study, and adjust accordingly (ID 10). Reflecting on this, a participant noted:

> "*So I know that the things are like the mixing a lot of things, but this is just for the reporting the results, the TRIPOD is just, I think, the tip of the iceberg. It is just, it is not all things (about reporting CPM study results)*." (ID 3)

#3-3-6 Selecting representative samples

Participants emphasized the need to align the sampled population with the broader group of intended applications, thereby protecting the model from potential bias and enhancing the reliability of the outcomes (IDs 7, 13, 14). They also warned against the risk of mistaking artificial subgroups created by non-representative samples for genuine samples of representative patients, highlighting the importance of distinguishing between them to preserve the model's real-world applicability (ID 18). For instance, a model developed using data from a specific demographic or

geographic area may not perform well when applied to a more diverse or different population, leading to biased or inaccurate predictions. Ensuring that the sample accurately reflects the intended population helps in generalizing the model's results and making reliable clinical decisions. Emphasizing this point, one participant noted:

> "*The population used to calculate the prediction model should fit to the population where you think you could apply this*." (ID 13)

#3-3-7 Using trustworthy datasets

Participants underscored those trustworthy results of a CPM study depends heavily on using datasets prepared according to responsible and accepted practice. They emphasized that it is not merely about having large datasets but ensuring they come from trustworthy sources and can withstand rigorous scrutiny (IDs 8, 10). Clarifying the standardization and definition of variables emerged as a non-negotiable aspect of establishing valid datasets (ID 11). Participants also encouraged the transparent standardization of definitions and meticulous validation of variables, especially from data sources like EMRs, ensuring consistent results as well as providing internal consistency values (IDs 8, 11). The other points of discussion also included the validation of data quality, focusing on the unequivocal clarity regarding the quality of data related to outcomes, predictor variables, and the validity of medical codes (IDs 20, 12). Summarizing these insights, one participant pointedly shared,

> "*I don't know anything about how good the data are that go into your model, right? You tell me that you identify these outcomes. I don't know whether those outcomes are validated or not. Most papers don't say how they validated the outcome*." (ID 20)

Stability

#3-4-1 Demonstrating model equity across individuals and populations

Participants underscored the imperative for reporting results of CPM studies to demonstrate stability, which extends beyond ensuring information remains consistent, and that model results are consistently delivered across populations. Emphasizing the importance of such model equity, they expressed concerns over prediction models that might inadvertently favor one group over another or provide misleading results when applied to diverse cohorts (IDs 7, 12). For example, a model that performs well for one demographic group but poorly for another could lead to unequal access to care and treatment outcomes. The risk of perpetuating health inequalities was a recurrent theme, particularly when models demonstrated skewed performance for certain groups (ID 7). Ensuring that models are tested and validated across diverse populations can help identify and mitigate biases, leading to more equitable healthcare solutions. As one participant eloquently stated:

> "*If prediction models work better in white people and work very bad in another group, then it might create inequalities*." (ID 7)

5.5 DISCUSSION

This study, conducted through interviews with 18 participants, identified needs among biomedical researchers across three themes—'understandable,' 'useful,' and 'trustworthy'—as quality attributes of CPM study results. All sub-themes from the three attributes are covered, with each sub-theme having at least one need, and some having up to seven needs.

Following those study findings, I will discuss 1) the Alignment of identified needs for reporting CPM study results of preimpact analysis studies with TRIPOD, 2) Overlapping themes,

3) Identified needs that warrant expanding the current reporting guidelines for CPM study results, 4) Addressing challenges by fulfilling needs, and 5) future studies that can be carried out.

Alignment of the needs with items for reporting preimpact analysis studies

Among all identified needs, many align with the reporting guidelines for common items to be reported in CPM study results of preimpact analysis studies outlined in references that guide researchers to conduct and report preimpact analysis studies (Collins et al., 2015; Steyerberg and Vergouwe, 2014). For example, in the 'understandable' theme under the 'analysis' sub-theme, the sequence of analysis aligns with model development in CPMs suggested, such as breaking down data exploration (#1-3-1), detailing steps of model development (#1-3-2) and reporting multiple metrics of model performance (#1-3-3). Furthermore, participants also explicitly supported adherence to TRIPOD as the main guideline for reporting preimpact analysis studies. These needs fall under the 'understandable' and 'trustworthy' themes. In the 'understandable' theme, under the 'knowledge' sub-theme, one need emphasizes providing basic information about study results uniformly, adhering to standard practices and guidelines, ensuring consistency and transparency (#1-1-1). Under the 'trustworthy' theme, within the 'validity' sub-theme, there is a need for reporting study results that adhere to standard practices (#3-3-1).

This alignment is expected, as all interview participants have experience as authors or reviewers of preimpact analysis studies, with no reporting experience in impact analysis studies. Despite the freedom to mention needs related to other types of CPM studies during interviews, only one need distinctly focused on impact analysis studies emerged under the theme 'useful,' specifically the sub-theme 'quality increase.' This need suggests that prediction models should demonstrate patient outcome improvements, which are achievable only through randomized controlled trials (RCTs) (#2-4-1). It is also noteworthy that two needs explicitly supported aspects

of TRIPOD within the 'understandable' and 'trustworthy' themes. In the 'understandable' theme under 'knowledge' sub-theme, one need to emphasize providing basic information about study results uniformly, adhering to standard practices and guidelines, and ensuring their consistency and transparency (#1-1-1). In the 'trustworthy' theme, under the 'objective' sub-theme, there was a call for reporting study results in adherence to standard practices (#3-3-1). Both imply support for utilizing standard practices as outlined in guidelines like TRIPOD.

Overlapping themes

The overlapping themes identified in this study not only resonate with the items suggested in TRIPOD, which focuses on transparency as a quality attribute but also show overlaps between the themes themselves. Figure 5-1 illustrates these overlaps, especially between the 'understandable' and 'useful' themes, as well as between the 'understandable' and 'trustworthy' themes. The approach to addressing overlapping themes in qualitative research can vary depending on whether the study uses an inductive or deductive analysis approach. In inductive analysis, such as thematic analysis, where codes/themes are generated during the analysis, overlapping themes are often collapsed to ensure distinct themes (Antes et al., 2019; Braun and Clarke, 2006). In deductive practices, where codes/themes are predetermined using existing frameworks, overlapping themes can be maintained as they are (Lester, 2022). Overlapping themes are also common when detailing concepts within different constructs (Campbell et al., 2013).

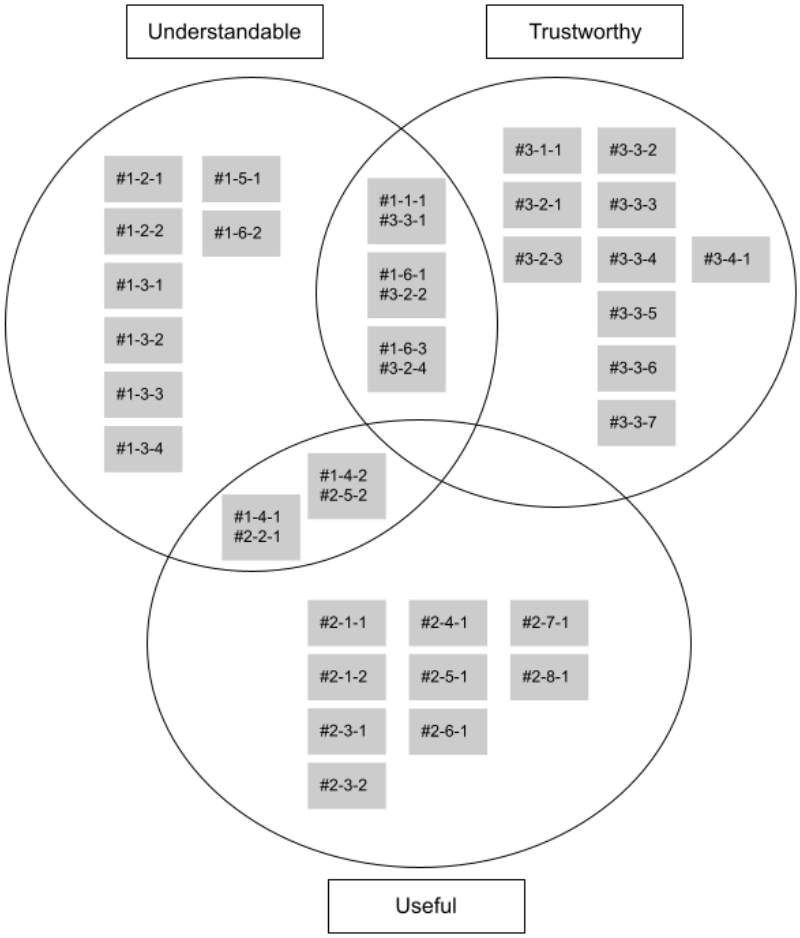

Figure 5-1 Unique and overlapping needs

In my study, these overlaps may occur for two reasons. First, overlaps for needs are found under definitions that share similarities, as shown in Table 5.1. For example, there is a notable overlap between the 'understandable' theme and its 'application' sub-theme, which closely relates to the overall concept of the 'useful' theme, particularly in demonstrating the applicability of CPMs. Furthermore, the definition of the 'evaluation' sub-theme under 'understandable' closely aligns with the 'trustworthy' theme, especially its 'objectivity' sub-theme, which pertains to evaluating the information presented in CPM study results. The second reason for overlaps is that the interviews were conducted in three distinct sections. Participants were asked about 'understandable' first,

'useful' after 'understandable,' and 'trustworthy' after 'useful' was complete. This sequence was mirrored in the analysis, where I maintained identified needs within each theme, even when aware of similar needs in other themes.

The implications of the observed overlaps in this study, especially between the 'understandable' theme and the 'useful' and 'trustworthy' themes, highlight an important interconnection among these quality attributes when reporting CPM study results. Specifically, enhancing the 'understandable' aspect of CPM study reports seems to concurrently address certain needs within the 'useful' and 'trustworthy' themes. This suggests that focusing on making study results more understandable could inherently contribute to their perceived utility and reliability. However, it is important to note that not all needs within the 'useful' and 'trustworthy' themes are covered by 'understandable.' Conversely, concentrating solely on 'useful' or 'trustworthy' may not inherently improve the other quality attributes. Therefore, efforts to ensure the quality of CPM study results across all three quality attributes should not be pursued in isolation.

Identified needs that suggest gaps in the current reporting guidelines for preimpact analysis studies, TRIPOD

Despite the alignment of the identified needs in this study with the reporting guidelines for preimpact analysis studies in TRIPOD, these needs also suggest that the current guidelines might need to keep pace with advancements in CPMs. As highlighted under the 'trustworthy' theme, sub-theme 'validity,' participants suggest adjusting to the limitations of established guidelines (#3-3-5), with adjustments reflecting advancements in the CPM field, as noted across all three themes. The following paragraphs highlight needs that reflect gaps in TRIPOD that were identified in my study.

In the 'understandable' theme, a notable need for authors to consider is designing presentations of CPMs that allow users to try out models through clear operational guidance, user-friendly scoring systems, and web-based applications for active testing and implementation in clinical settings (#1-4-1). Ensuring presentations meet these requirements likely requires involving clinicians, the target users of CPMs in clinical settings. Such designs could benefit from the Human-Centered Design (HCD) framework to guide authors (FDIs, 2009). Unfortunately, current CPM studies focusing on preimpact analysis studies have not actively involved clinicians in designing these presentations. Studies that do involve target users often focus solely on designing CPMs, separate from preimpact analysis studies (Chiang et al., 2021).

In the 'useful' theme, an important need cautions authors about presenting CPM results. Needs (#2-1-1) and (#2-1-2), under the 'critical support' sub-theme, emphasize guiding the use or non-use of prediction models by clearly defining their intended functions and outlining potential limitations and risks to prevent misuse. This consideration is particularly critical for skeptical readers, especially when authors might overestimate the models' capabilities, leading readers to seek ways to avoid overreliance on these models (Karhade and Schwab, 2020). Thoroughly addressing these aspects in research papers not only ensures that CPM study results are trustworthy but also promotes a more informed application of CPMs in clinical practice, ensuring they support rather than replace clinical judgment (Sezgin, 2023; Waljee et al., 2014).

In the 'trustworthy' theme, a notable need under the sub-theme 'stability' concerns demonstrating model equity across individuals and populations (#3-4-1). Whether a CPM discriminates against population subgroups is a significant issue in the field (Obermeyer et al., 2019). Authors should address this by reporting any potential biases during model development that would cause this issue. Another concern under 'trustworthy' is model accuracy. As model

accuracy is a primary goal of CPM studies (Shipe et al., 2019), participants understand that authors will aim for this. However, there is concern that achieving accuracy might come at a cost. Under need (#3-2-3), participants emphasize the need for rigorous reporting and justification of all decisions made in CPM study research processes to convince readers of the justifiability of claimed model performance levels.

Addressing challenges by fulfilling the needs

This section focuses on discussing biomedical researchers' needs that may help address challenges they experience regarding understandable, useful, and trustworthy results in CPM identified in Chapter 4. By dissecting these needs under the key themes of 'understandable,' 'useful,' and 'trustworthy,' this discussion offers a clearer understanding following my study finding on Chapter 4 of how to address the inherent challenges to ensure that results of CPM studies contain those three quality attributes.

For results to be understandable, Chapter 4 identified challenges in CPM often arising from the complex interaction between model intricacy and user background. This complexity includes difficulties in interpreting statistical nuances and conveying intricate concepts to a broader audience, such as healthcare providers. In this chapter, the needs identified under the 'understandable' theme by participants provide suggestions to help overcome these challenges, particularly under the sub-themes of 'comprehension' and 'application'. For 'comprehension,' participants suggested ensuring user-centric reporting. They called for making study results understandable to a broad audience, including those with limited technical expertise, notably clinicians as the target users (#Needs: 1-2-1). This could involve including target users, such as clinicians and statisticians, in author teams to bring diverse expert perspectives. For 'application,' participants recommended allowing users to interact with prediction models through clear

operational guidance (#Needs: 1-4-1). This involves user-friendly scoring systems and the development of web-based applications for active testing and implementation in clinical settings, enhancing practical applicability, and user engagement as well as promoting better understanding of intricate concepts to the broader audience on how the CPMs work.

In the context of useful results, Chapter 4 highlighted the multifaceted challenges of making prediction models clinically relevant and adaptable across varied clinical practice settings. This study underscores the needs within the 'useful' theme, particularly on the sub-themes of ‘critical support,’ ‘easy to use,’ and ‘performance increase.’ Under the ‘critical support’ sub-theme, the needs identified include guiding the use of prediction models. This involves clarifying their intended functions, outlining potential limitations, and highlighting risks to prevent misuse. Additionally, it includes cautioning against the unintended consequences of inappropriate model use, such as potential pitfalls, ethical concerns, healthcare burdens, and the possibility of treatment deterrence due to model predictions (#Needs: 2-1-1 & 2-1-2). For the ‘easy to use’ sub-theme, the need emphasized is making prediction models user-friendly (#Needs: 2-2-1). This involves portraying these models as straightforward and accessible, typically through intuitive tools and interactive digital platforms, to convince users of their ease of use. Lastly, under the ‘performance increase’ sub-theme, the focus is on demonstrating how prediction models can enhance clinicians' performances (#Needs: 2-6-1). This need revolves around facilitating better-informed decisions and seamlessly integrating these models into clinical workflows, thus contributing to improved clinical practices.

Ensuring trustworthy results in CPM studies poses a complex challenge, as discussed in Chapter 4. This challenge encompasses limited access to original data and codes, difficulties in replicating experiments, and issues with handling missing data. Additionally, rigorous data quality

and methodological checks, as well as consistent reporting, are essential to ensure trustworthy results. This study identifies specific participant needs to address these challenges, particularly under the sub-themes of 'objective' and 'validity'. Under 'objective', participants emphasize the need for reproducibility, transparency, open science frameworks, online code sharing, and structured data-sharing mechanisms (#Needs: 3-2-2). They also highlight the importance of justifying research decisions, selecting analytical methods and variables carefully, addressing anomalies, providing detailed methodological disclosure, and handling missing data appropriately (#Needs: 3-2-3). In the 'validity' sub-theme, participants advocate for adherence to study methods and goals (#Needs: 3-3-3), using representative populations (#Needs: 3-3-6), and using trustworthy datasets with standardized variables (#Needs: 3-3-7). These comprehensive suggestions aim to address the multifaceted challenges of ensuring trustworthy results, rigorous data quality, and consistent reporting in CPM studies.

This study reinforces and expands on the findings from Chapter 4, highlighting a consistent pattern where addressing the challenges associated with achieving 'useful' and 'trustworthy' results in CPM studies may be relatively more demanding compared to 'understandable' results. Thus, this study also reflects that while ensuring understandable results of CPM study may be challenging among many authors, especially those with less experience as identified in Chapter 4, the steps towards achieving useful and trustworthy results of CPM studies involve a heightened level of complexity and thoroughness. Furthermore, this study demonstrates that identifying challenges in attaining understandable, useful, and trustworthy results in CPM studies and addressing them by identifying specific needs is feasible. The semi-structured interviews conducted here provide a deeper understanding of these challenges and needs, enhancing the insights previously outlined in

Chapter 4 and enriching our comprehension of what constitutes understandable, useful, and trustworthy results in CPM studies.

Limitations

This study, while providing valuable insights into the needs associated with reporting understandable, useful, and trustworthy CPM study results, is subject to several limitations inherent in its design and methodology.

1. Focus on CPM Study Results: This study is limited by its exclusive focus on the results section of CPM studies, potentially overlooking the interconnectedness of other sections, such as the methodology or introduction and their impact on the overall quality and comprehensibility of CPM study reports. By concentrating solely on the results, discussions, and conclusions, the study may miss crucial elements in other sections that significantly contribute to the readers' understanding and evaluation of CPM studies. This focused approach may not fully capture the holistic reporting practices necessary for comprehensive and effective communication of CPM research findings.

2. Representation of Samples: The representativeness of the sample in this study presents another limitation, as participants were drawn from a specific subset of survey respondents who expressed willingness to participate in follow-up interviews. This self-selection process might lead to a biased sample that does not fully represent the diversity of perspectives among all authors and reviewers of CPM research papers. The limited number of interviewees further restricts the generalizability of the findings, as the experiences and views of a broader group of biomedical researchers in the field of CPM might vary significantly from those captured in this study.

3. Dependence on self-report: The study's reliance on participants' self-reported experiences and perceptions introduces a limitation related to potential recall bias and subjective interpretations. Participants' accounts of their experiences with writing and reviewing CPM study results might be influenced by their memory, personal biases, or desire to present themselves in a certain light. This dependence on self-reported data may affect the accuracy and objectivity of the insights gathered, thereby influencing the study's conclusions about the need to report CPM study results.

4. Deductive Approach with Predetermined Themes: Employing a deductive approach with predetermined themes and sub-themes based on existing frameworks restricts the study's ability to uncover novel insights or emerging themes outside the established codebook. While this approach ensures a structured analysis aligned with the study's objectives, it may limit the exploration of unanticipated areas relevant to the quality attributes of CPM study results. Even though I tried to inquire about additional themes from participants, no entirely new themes emerged from the participants' suggestions that warranted the creation of additional themes. Notably, themes like reproducibility and replicability were often mentioned when I inquired about additional quality attributes beyond ‘understandable,’ ‘useful,’ and ‘trustworthy,’ but these were aptly included under the existing 'understandable' and 'trustworthy' themes.

5. Coding Process: The coding process in this study, involving two coders and aimed at achieving a high level of inter-coder reliability, presents limitations related to the subjective interpretation of qualitative data. While achieving a Cohen's Kappa score of 0.8 indicates a high level of inter-coder reliability, the initial coding discrepancies, and the iterative process to resolve these discrepancies may reflect challenges in applying the codebook consistently across diverse participant responses. This aspect of the methodology could influence the subsequent

interpretation of needs, as the resolution of coding discrepancies is subject to the coders' judgment and consensus.

Future studies

This Chapter identified the needs of biomedical researchers to present the results of CPM studies in three quality attributes: understandable, useful, and trustworthy, from the perspectives of biomedical researchers as authors and reviewers. Further investigations should involve clinicians as participants in a study similar to mine because, in the present study, my participants acknowledged clinicians also as a primary target audience for CPM research papers. These future studies can inquire into the needs among clinicians to ensure that CPM study results are understandable, useful, and trustworthy.

Additional exploration might include developing quantitative assessment metrics to evaluate the quality of CPM study results. Then, these metrics could be used in studies that involve comparing the different qualities of CPM study results. This comparison could aim to understand how these qualities may influence users' perception of information, and whether the study results are understandable, useful, and trustworthy for the target audiences, either biomedical researchers or clinicians. Furthermore, whether studies that design the reports of CPM studies to fulfill the needs identified in this study would improve the perception of biomedical researchers remains to be seen. Such studies may employ a comprehensive human-centered design (HCD) approach (FDIs, 2009).

5.6 CONCLUSION

In conclusion, this study has provided a comprehensive and deeper insight into biomedical researchers' needs to ensure understandable, useful, and trustworthy results of CPM studies—

perspectives that are currently lacking in the well-known guidelines for reporting CPM studies. In this study, I identified a diverse array of needs spanning across three key themes: 'understandable' with 13 needs across 6 sub-themes, 'useful' encompassing 11 needs within 8 sub-themes, and 'trustworthy' comprising 13 needs spread over 4 sub-themes. These needs predominantly align with reporting for preimpact analysis studies of CPMs. I discovered needs that specifically support adherence to TRIPOD guidelines for reporting CPM study results. Additionally, I found the need that suggest authors to comply with advancements in the CPM field and identify gaps in TRIPOD, such as accompanying presentations of CPMs using user-friendly tools for their application, detailing when to use CPMs, and convincing readers that CPMs ensure model equities across subgroup populations. While most needs are unique across the three quality attributes, overlaps in certain needs were observed, particularly with 'understandable' frequently overlapping with both 'useful' and 'trustworthy.' These overlaps provide further insights that achieving one quality attribute ('understandable') may aid in achieving the others, but not all, necessitating meticulous efforts to consider needs across these three quality attributes.

Chapter 6. IDENTIFYING VISUALIZATION PREFERENCES AMONG BIOMEDICAL RESEARCHERS FOR UNDERSTANDABLE, USEFUL AND TRUSTWORTHY RESULTS IN CLINICAL PREDICTION MODEL STUDIES: AN INTERVIEW STUDY

6.1 INTRODUCTION

CPMs provide risk estimates for the presence of disease (i.e., **diagnosis**) or an event in the future course of disease (i.e., **prognosis**) in individual patients (Steyerberg and Vergouwe, 2014). However, digesting information from CPMs and incorporating them into decision-making is often challenging for its target users, who are primarily clinicians (Kappen et al., 2016; Walsh et al., 2021). These challenges stem from the complexity of prediction models, the probabilistic nature of outcomes, and the lack of corresponding management recommendations (Kappen et al., 2016). This issue can lead to negative consequences in their implementation and erode trust among the target users (Walsh et al., 2021). Furthermore, as reported in Chapter 4, my study also identified that these challenges persist among biomedical researchers who produce these CPMs.

Visualizations serve as a powerful tool to communicate CPM to their target users, enabling the delivery of contextualized information (Cui, 2019; Keim et al., 2008; Walsh et al., 2021). These visual tools can translate intricate mathematical models as well as risk and probabilities of CPMs into accessible representations (Gotz and Borland, 2016). Incorporating visualizations into the presentation of prediction models has become prevalent (Bonnett et al., 2019; Chiang et al., 2021; Van Belle and Van Calster, 2015). As introduced in Chapter 1, understandable, useful, and trustworthy visualizations are key quality attributes focused on in previous studies in visualizations

for presenting prediction models (Burns et al., 2020; Chatzimparmpas et al., 2020; Reeder et al., 2011).

TRIPOD also recommends the use of visualizations that can accompany the reporting of results of preimpact analysis studies, providing examples that authors can use (Moons et al., 2015). However, these visualizations do not encompass all six stages of developing or validating prediction models, including data exploration, predictor selection, modeling, result exploration, model performance, and model presentation, as introduced in prior work in Chapter 1. Furthermore, whether these visualizations improve the quality of CPM study results in terms of the three quality attributes—understandable, useful, and trustworthy—remains unknown. Thus, a study that identifies visualization preferences for presenting understandable, useful, and trustworthy CPM study results among biomedical researchers as the target users remain absent.

Application of Munzner's Nested Model on presenting visualizations of CPM study results

Munzner's Nested Model suggests a four-level guide to design visualizations (Munzner, 2009). These levels encompass domain problem characterization, data/operation abstraction design, encoding/interaction technique design, and algorithm design. Authors of CPM research papers could apply this framework by collaborating with domain experts or consulting existing literature references (Kerracher and Kennedy, 2017).

The first level (Level 1), domain problem characterization, involves identifying the data to be visualized and the target audiences' need to use the visualization. At this stage, authors of CPM research papers start characterizing target audiences' needs for involving visualizations to present the complexities of the results from developing or validating CPM. This initial step is critical in ensuring that the visualization aligns with the goals of the research and addresses the specific challenges or questions that readers may encounter when interpreting CPM study results.

The second level (Level 2), data/operation abstraction design, is crucial in transforming domain-specific data into formats suitable for visualization and abstracting operations to operationalize tasks based on readers' needs identified in the first level, referred to as visualization tasks for this study. This stage holds significant importance for authors of CPM research papers. It involves the essential task of translating results from developing or validating CPMs into visualizations. Moreover, this stage requires defining the visualization tasks and identifying potential complexities in the visualizations.

The third level (Level 3), encoding/interaction technique design, advances the design of visual encoding and interaction techniques. At this stage, authors determine the visual encoding that best represents the visualization tasks for their target audience. If the decision is made to incorporate interactivity into the visual presentation, designing user interactions becomes a critical component. Regardless of whether the visualizations are interactive, they should strike a balance between visual appeal and functionality.

The fourth level (Level 4), algorithm design, involves tasks that authors undertake when deciding to create interactive visualizations for presenting CPM study results in research papers. This level may focus on designing a prototype of interactive visualizations. The design process involves selecting appropriate algorithms for presenting interactive visualizations based on the visual encodings identified in the third level, optimizing performance for interactivity, and developing code for these visualizations. Rigorous usability testing is crucial to verify the correctness and efficiency of the algorithms, ensuring that the visualizations are user-friendly, responsive, and effective. However, most authors may opt not to pursue this due to the potential complexity of the presentation in research papers. Consequently, this study does not extend to this level.

6.2 STUDY OBJECTIVES

Following Munzner's Nested Model, the objective of this study is to identify the needs for visualizations (Level 1), preferred visualization tasks (Level 2), and visual encoding techniques (Level 3) in presenting CPM study results among biomedical researchers to ensure the results of CPM studies are understandable, useful, and trustworthy. Thus, I framed the research inquiry into three research questions:

RQ1: How do biomedical researchers characterize the need for using visualizations among their target audience?

RQ2: What visualization tasks and encodings are preferred by biomedical researchers in the presentation of results of CPM studies?

RQ3: How can visualizations preferred by biomedical researchers ensure understandable, useful, and trustworthy results in CPM studies?

6.3 METHODS

Recruitment and data collection, and ethics

Recruitment, data collection, and ethics approval are described in Chapter 5.

Data Analysis Process

The qualitative data analysis for this chapter also employs directed content analysis and uses the paragraph chunks from the spreadsheets that I produced in Chapter 5. The coding process also follows the procedure referred to as the Inter-Coder Reliability (ICR) workflow with two coders (O'Connor and Joffe, 2020). Coder 1 (IR) initiated the coding process by separating the paragraph chunks that mentioned topics related to visualization preferences, visualization tasks, and visual encoding during the analysis in Chapter 5 to answer RQ1 and RQ2 of this Chapter.

To answer RQ1, Coder 1 created two codes related to preferences in using visualizations (Munzner Level 1): 'using visualizations' and 'accompanying visualization with text'. These codes aim to describe biomedical researchers' preferences in using visualizations to report CPM study results. These codes represent responses that discuss the general use of visualizations and text to present CPM study results without specifically emphasizing whether these preferences ensure that CPM study results are understandable, useful, and trustworthy. Therefore, codes for the three themes representing the three quality attributes were removed from the spreadsheet. Thus, the final spreadsheet for analyzing RQ1 contains columns for ID, text, and coding for preferences in using visualizations.

For RQ2, the coders selected paragraph chunks and assigned codes representing visualization tasks expressed by participants for presenting results in CPM studies. Each code corresponds to one theme for a visualization task (Munzner Level 2). Coder 1 categorized these visualization tasks into six sections of result presentation in CPM studies: data exploration, feature selection, modeling, result exploration, model performance, and model presentations as introduced in prior work in Chapter 1 (Lu, 2017; Steyerberg and Vergouwe, 2014). Coder 1 also devised a coding scheme for the identified visualization tasks for Coder 2 to apply. Each visualization task was further divided into more detailed sub-tasks. For each sub-task, visual encodings (Munzner Level 3) were provided. After identifying the visualization tasks, the next was determining visual encodings for each sub-task. These visual encodings were mostly derived from participants' responses. However, when participants mentioned visualization tasks without specifying visual encodings, the encodings were taken based on literature or best practices. At other times, when participants provided their expected visualization encodings, the visualization tasks or sub-tasks were described based on those encodings. The final spreadsheet as seen in Appendix I for this

analysis contains columns for ID, text, visualization tasks, visual encoding, and themes from Chapter 5.

Once Coder 1 finished the coding process for RQ1 and RQ2, Coder 1 removed the codes column to create a 'clean' file, leaving only the paragraph chunks for Coder 2 to assign codes. Each spreadsheet, one for RQ1 and one for RQ2, was prepared for Coder 2's work. The spreadsheet for RQ1 that Coder 2 worked on contained columns for ID, text ('paragraph chunk'), and preferences in using visualization. The spreadsheet version Coder 2 used for RQ2 included the following columns: ID, text, and visualization tasks. Coder 2 then assigned codes for preferences in using visualization and visualization tasks to the paragraph chunks in the text column for the spreadsheets of each RQ1 and RQ2.

From this point, the process of achieving ICR was the same as in Chapter 5, where Coder 2 started with the first two transcripts from two participants and then paused to discuss any discrepancies with Coder 1. This iterative process continued, with Coder 2 (KN) completing one more transcript and matching the coding with Coder 1, until a Cohen's kappa score of at least 0.8 was achieved between the two coders. After coding was completed for RQ1 and RQ2, Coder 1 used R to assess our ICR. Upon completing the coding process, Coder 1 extracted information from each code to answer the RQ1 and RQ2.

For RQ3, Coder 1 cross-referenced paragraph chunks and themes for visualization tasks from RQ2 with the themes from Chapter 5, which consist of three quality attributes of reporting results in CPM studies, 'understandable,' 'useful,' and 'trustworthy.' Therefore, each paragraph chunk about visualization tasks has additional codes about participants' needs for quality results in CPM studies across those three quality attributes. By cross-referencing both sets of codes, Coder

1 determined how specific visualization tasks could contribute to achieving results in CPM studies that are more understandable, useful, and trustworthy.

6.4 RESULTS

Participant characteristics

The participant characteristics are described in Tables 5.1 and 5.2 of Chapter 5. However, one participant did not provide answers about the use of visualizations in reporting the results of CPM studies.

6.4.1 *RQ1: How do biomedical researchers characterize the need for using visualizations among their target audience?*

I identified two themes that answer this research question: using visualizations and accompanying visualizations with text to report the results of CPM studies. Thirteen participants expressed their opinions on these themes. Appendix H provides the complete answers from the participants.

Using visualizations to report results of CPM studies

Reflecting upon the role of visuals in effective communication, participants emphasized that visualization is an integral part of reporting CPM studies (IDs 2, 3, 5, 7, 8, 10, 12, 19). They further underscored the role of tables, figures, and other visual aids in efficiently presenting complex data sets and findings to their target audiences (IDs 7, 8, 17). Participants also recognized that visualizations could expose nuances and unique elements within the data that might be overlooked in textual descriptions alone and eventually improve the reporting quality of results of CPM studies (IDs 3, 17). As one participant clarified,

"*And then also maybe having some mandatory figures that if the journal mandates that I need to look at. ... So with this, if they made a section in the results for explainability or understandability, I think that would improve the quality of our papers a whole lot more.*" (ID 17)

Furthermore, participants noted that visually appealing figures often succeeded in attracting reader attention and retaining engagement (IDs 10, 18). The use of distinct visual elements such as color or multi-layered information was seen as not only improving reader comprehension but also enhancing the aesthetic appeal of the work (ID 10). One participant suggested:

"*Maybe you have some way of putting colors or making it more fancy, maybe combine different information in one figure or in different panels. I do think that will seriously boost your chances of getting into better journals or just getting people more engaged with your paper*." (ID 18)

Yet, alongside the benefits of utilizing visuals, participants also cautioned against their overuse or misrepresentation (IDs 2, 6, 8). While visualizations offer critical support to the understanding of complex data, excessive complexity in visualizations could cloud the intended message, potentially leading to reader confusion (ID 6). Misleading practices, such as changing the dimensions of a graph to exaggerate differences, were warned against (ID 8). As one participant suggested,

"*But doing things like changing the dimensions of a chart or a graph to make it seem as if the change that you're seeing in a variable or the difference between groups or something is much larger than what it is.*" (ID 8)

Accompanying visualization with text to report results of CPM studies

Participants emphasized that readers should not be left in the dark when interpreting visual data. They suggested providing context through text when presenting visual data (IDs 17, 18). This textual accompaniment ensures that the results or the intended message are not misconstrued. This text should not be limited to figure legends or table captions only but should further set the context for the presented visualization. A notable statement from the discussions was:

> "*I think the most important piece is that the text that you're writing, for me as a reader, what I see visually needs to be set in context in the text as well.*" (ID 17)

Furthermore, participants suggested that accompanying text in graphical formats should be balanced to allow for a more comprehensive understanding (IDs 8, 18, 20). While the text can pinpoint and summarize crucial points, visuals offer more detailed insight, catering to readers who might find graphical data more intuitive (IDs 8, 19, 7). As one participant noted,

> "*I think, I think both. Both. I think the visualization should, how to say, visualization should allow us to imagine the product and its value. But of course, it is hard to summarize everything into just one graph or one figure. So, in that sense, the text is always nice to supplement things. But the text should not be the first thing to look at, because it is often very heavy.*" (ID 7)

6.4.2 *RQ2: What visualization tasks and encodings are preferred by biomedical researchers in the presentation of results of CPM studies?*

The following analysis explored participants' detailed preferences for visualizations designed to make the results of CPM studies understandable, useful, and trustworthy. A total of 16 participants out of 18 contributed to the identification of visualization tasks and visual encodings, which I distributed into the six sections of presenting results in CPM studies (Lu, 2017),

with most emphasizing model presentations (10 participants), while predictor selection and incorporating confidence intervals received the least attention (2 participants each). A total of 28 identified visualizations (task/sub-task and encoding) were noted. At least one visualization task was identified for each section, with up to three visualization tasks identified in result exploration. For each visualization task, there was at least one visualization sub-task/encoding, with up to four visualization sub-tasks/encodings in result exploration, making it the section with the most tasks, sub-tasks, and encodings. The following report provides the details of visualization tasks and visual encodings for each section in presenting CPM study results. Furthermore, Appendix I provides the complete participant answers used to analyze this research question.

6.4.2.1 Data exploration

Data exploration, commonly known as data inspection or preprocessing, is a foundational stage in developing a CPM. This phase usually provides the foundation for further analysis in the prediction model development. Our analysis found a group of visualization tasks and encodings related to data exploration focused on detecting initial data patterns, trends, or anomalies, particularly outliers. Table 6.1 below summarizes visualization tasks and their sub-tasks, followed by the corresponding visual encodings. Similar tables will also be provided for the other sections of the result presentation in CPM studies. Numbering for visual encodings is also included in the table for better navigation to each visualization of interest.

Table 6.1 Identified visualization tasks for the data exploration section

Visualization tasks - subtasks	Visual encoding	#Visual encoding
Detecting initial data patterns, trends, anomalies, and outliers (IDs 3, 14, 17, 19)		
Displaying the distribution of a variable	Scatter plots that allow for the detection of non-linearity Outlier detection plots such as boxplots or violin plots.	1-1

Visualization tasks - subtasks	Visual encoding	#Visual encoding
Showing detailed data properties	Histogram for feature characteristics Box plots for ceiling or floor effects	1-2
Displaying the distribution of raw data, transformed data, and the effect of data transformations.	Plots such as histograms, Q-Q plots, Boxplots, violin plots, line plots, or scatter plots that show normality distributions before and after transformation.	1-3

Detecting initial data patterns, trends, or anomalies, outlier

Visualization task: Participants preferred presenting data exploration results in CPM studies, focusing on *detecting initial data patterns, trends, anomalies, and outliers.* The first sub-task emphasizes *displaying the distribution of a variable*, as visualizations can effectively represent a clear snapshot of data distribution, which is essential for further analysis. (IDs 3, 19). Furthermore, *showing detailed data properties i*s vital as it allows for a deeper understanding of the characteristics of each feature, as well as any ceiling or floor effects (IDs 17, 14). In addition, *displaying the distribution of raw data, transformed data, and the effect of data transformations* is a significant step in the data discovery phase, ensuring the data meets the necessary assumptions before proceeding to model building and also aiding in the interpretation of model results (ID 17). This meticulous approach towards visualization tasks in presenting results not only enhances the clarity and comprehensibility of the findings but also invites a thorough discussion on the data characteristics for the model development that are often excluded in presenting results in CPM studies (ID 3). As one participant succinctly noted,

> "*First of all, to show the readers what signals the data may have in the beginning, just to show what kind of characteristics the data has, how much variability there is there in each of the features that's being questioned… Those kind of things, just to first of all present the*

data. Data visualization can go a long way. And again, this is not so much explored in the results section of papers." (ID 17)

Visual encodings. Visualizing these tasks in presenting the results of CPM studies opens up a window of exploration and understanding fundamental for both the users (Bruce and Bruce, 2017). *Scatter plots,* for instance, are instrumental in detecting non-linearity, thus laying down a foundational understanding of the relationships within the data. *Outlier detection plots* like boxplots or violin plots are essential in identifying data points that deviate significantly from the others, which could be critical in understanding and addressing issues related to data quality or underlying phenomena. Diving deeper into feature characteristics, *histograms* provide a lens through which the central tendency and spread of a particular feature, thereby providing a thorough understanding of each variable's distribution. *Box plots* stand out as a robust tool for identifying ceiling or floor effects, highlighting the range and dispersion of data, which is crucial in understanding the limitations and the scope of the analysis (Arslan and Benke, 2023). Various plots can be utilized to display the distribution of raw data, transformed data, and the effects of data transformations. Histograms, Q-Q plots, boxplots, violin plots, line plots, or scatter plots are particularly effective. These visualizations demonstrate the normality of data distributions before and after transformation, illustrating the changes in distributions and ensuring that the data aligns with the assumptions necessary for subsequent modeling. This range of visualization tasks captures the core aspects of data exploration in model development.

6.4.2.2 Predictor selection

Predictor Selection, also known as Model Specification, is the step where relevant features are identified. Reporting this section may demonstrate a careful selection of predictors, considering various strategies and avoiding common pitfalls such as overfitting (Steyerberg and Vergouwe,

2014). My analysis identifies one visualization task preferred by participants: visualizing predictor selection. Table 6.2 summarizes the visualization task, sub-tasks, and feature selection encoding.

Table 6.2 Identified visualization tasks in the predictor selection section

Visualization tasks - subtasks	Visual encoding	#Visual encoding
Visualizing predictor selection (IDs 17, 18)		
Analyzing feature independence	Correlation matrix heatmaps	2-1
Showcasing predictor selection and their effects	Tables with predictors, coefficients, and intercepts	2-2
Depicting the process of predictor selection	Flowcharts or process diagrams	2-3

Visualizing predictor selection

Visualization tasks. Participants discussed their preferences in presenting results of CPM studies focusing on *visualizing feature selection for developing the models. Analyzing feature independence* is one sub-task in this group that assists in identifying multicollinearity among predictors in a linear regression model. It is one important step in ensuring the robustness and reliability of the model, especially when using linear regression (ID 17). Furthermore, the task of *showcasing predictor selection and their effects* is also vital as it clarifies the final predictors chosen for the model along with their corresponding coefficients and intercepts, offering a clear understanding of how each predictor contributes to the model (ID 18). Additionally, depicting the process of predictor selection step-by-step, including the criteria for inclusion or exclusion at each stage, highlights the model-building process (ID 18). This meticulous approach to visualizing predictor selection would encapsulate the importance of a structured visual narrative in understanding the evolution of predictor selection. As one participant aptly highlighted,

"*I think the paper, or when you report the risk prediction model, it is like a cascade or a story. So, I started with 100. I eliminated these because of missing data or whatever. I*

eliminated these because they were not associated with the outcome or whatever, and then I eliminated these because..." (ID 18)

Visual encoding. To visualize the tasks for feature selection in presenting results of CPM studies, authors can employ various visualizations (Lu, 2017). *Correlation matrix heatmaps*, for instance, provide an illustration of the degree of correlation among pairs of predictors for identifying multicollinearity. These heatmaps, with their color-coded cells, offer a visual simplicity that facilitates a quick understanding of the relationships between predictors, aiding in the analysis of feature independence. On the other hand, *tables with predictors, coefficients, and intercepts* serve as means to showcase predictor selection and their effects. Each row in such a table representing a predictor, along with columns for the coefficient and intercept, provides a clear and organized view of the final predictors chosen for the model. Color coding can further enhance the readability of the table, helping to distinguish between positive and negative coefficients, thus adding a layer of visual clarity to the representation. Furthermore, the depiction of the process of predictor selection can be visually captured using *flowcharts or process diagrams*. These graphical tools can illustrate the step-by-step reduction of the initial pool of predictors, with each node or step representing a decision point along with the associated criteria. Such visualizations encapsulate the systematic approach taken in predictor selection, providing a narrative that's easy to follow and understand, echoing the participant's sentiment about the paper reporting the feature selections of CPMs as a "cascade or a story."

6.4.2.3 Modeling

Modeling is the core stage where the predictive model is constructed and trained. It involves selecting an appropriate algorithm, defining the model's structure, and setting parameters (Lu, 2017). The following details how participants in CPM papers emphasize the importance of

explaining the model architecture and detailing the complex process and algorithm in the prediction model. This section only contains one visualization task, and no sub-tasks were found as depicted in Table 6.3.

Table 6.3 Identified visualization task for the modeling section

Visualization tasks - subtasks	**Visual encoding**	**#Visual encoding**
Explaining model architecture (IDs 6, 12, 17)	Decision trees or rule sets Model diagrams	3-1

Explaining model architecture

Visualization tasks. Participants showed a preference for presenting results in CPM studies with detailed visualizations that *explain the model's architecture* (ID 17). They emphasized the importance of visual representations that can effectively demonstrate the components of the models and the data flow from input to output (ID 17). These visualizations are crucial in conveying how the models are designed and built, as well as playing a significant role in the interpretation of results and findings (ID 17). The complexity often associated with black box models in machine learning and artificial intelligence necessitates clear and understandable visualizations to aid in explaining the methodology, particularly how variables are selected (ID 12). However, there is an acknowledgment of the challenges in fully visualizing all aspects of complex prediction models, especially those involving neural networks, where direct understandability might not always be achievable (ID 6). As one participant put it,

> "*I think it is important, we can use visual representation to our machine learning models not only in how they were designed and built, but also in the interpretation of the results and findings.*" (ID 17)

Visual encodings. In representing the task above in visualizations, authors may use *decision trees or, rule sets, and model diagrams* (Lu, 2017). If the model type supports it, *decision trees or rule sets* can be visualized to explicitly demonstrate how input variables are navigated to arrive at a prediction (Ozcan and Peker, 2023). This explicit illustration helps unravel the logic encapsulated in the model, making the decision-making process transparent to the readers. Furthermore, for more complex models, such as neural networks, *model diagrams* can showcase the layers, nodes, and connections (Liu et al., 2023). These diagrams provide a high-level view of the model structure, delineating the architecture and the interconnections that drive the model's predictions.

6.4.2.4 Result exploration

Result Exploration is an essential phase in the study of CPMs, focusing on the qualitative and visual analysis of modeling results (Lu, 2017). Various graphical representations are leveraged to dissect the model's behavior and performance, examining how variables influence the model output and highlighting the importance of individual predictors. This comprehensive exploration serves to identify patterns, trends, and potential areas for improvement while demonstrating the intricate relationships of predictors and their consequential impacts on model output. Acting as a bridge between the complex technical modeling process and practical application, this stage aids in translating multifaceted results into actionable insights. In this section, I found three visualization tasks: highlighting predictor importance in CPMs, demonstrating relationships of predictors and their impacts on model output, and assessing the models at different cut-offs or thresholds. Table 6.4 lists those visualization tasks, their corresponding sub-tasks, and visual encoding.

Table 6.4 Identified visualization tasks for result exploration section

Visualization tasks - subtasks	Visual encoding	#Visual encoding
Highlighting predictor importance in clinical prediction models (IDs 6, 10, 13, 19)		
Showing top predictors	Ordered list or bar chart	4-1
Showcasing top predictors by magnitude and effects (Harmful vs. Protective)	A forest plot	4-2
Exploring the effects of subsets of predictors on the model	Interaction plots	4-3
Demonstrating relationships of predictors and their impacts on model output (IDs 2, 6, 8, 17)		
Allowing users to demonstrate whether varying one predictor with all other predictors held constant would affect the model's risk predictions.	Tornado diagram	4-4
Showing users how varying levels of continuous variables influence associated hazard or odds ratios.	Restricted cubic splines	4-5
Manipulating multiple predictors for realistic scenarios	Interactive parallel coordinate plot	4-6
Assessing multiple interactions examines how combined variables affect the outcome for optimized model performance.	Interaction effect plot with variations	4-7
Assessing models at different cut-offs or thresholds (IDs 7, 11, 20)	Decision curve plot for net benefit.	4-8

Highlighting predictor importance in clinical prediction models

Visualization tasks. Participants expressed their preferences for highlighting predictor importance in presenting results in CPM studies (IDs 13, 6). The first sub-task involves *visualizing the most influential predictors of a model*, determined by the magnitude of their coefficients (ID 6). Another task is showcasing top *predictors by magnitudes, whether they have harmful or protective effects on the models* (ID 6). Additionally, *exploring the effects of subsets of predictors*

on the model provides a more granular understanding of predictor influence and the robustness of model performance (IDs 6, 10). This exploration can reveal insights into how variations in predictor inclusion impact the overall model. As one participant aptly put it:

> "*I think having kind of a top five or top three, or something like that for maybe both harmful and protective effects, or, you know, maybe just top five by magnitude and say whether they're harmful or protective. I think that is a good way to maybe present at least some of your results and hopefully, you know, hopefully, there's something there that is immediately obvious.*" (ID 6)

Visual encoding. To effectively communicate the tasks identified by participants, a range of visual encodings can be utilized. For showing the top predictors, *an ordered list or bar chart* serves as an efficient means to convey the hierarchy of importance among predictors (Devanarayan et al., 2023). This straightforward representation offers a clear visualization of predictor rankings. When it comes to displaying the top predictors by magnitude and effects (harmful vs. protective), an *ordered list or bar chart, enhanced with colors or symbols, can distinguish between harmful and protective effects*. This adds an additional layer of information, enriching the visual representation. A *forest plot* is particularly adept at illustrating the estimates of effect (like risk) for various predictors, complete with their confidence intervals (An et al., 2021). This allows for a comparative view of these factors, making the plot valuable for detailed analysis. In exploring the effects of subsets of predictors on the model, *interaction plots* can be effective. *Interaction plots* are instrumental in visualizing the interaction effects between subsets of predictors (Nahhas, n.d.). These plots contribute significantly to understanding the model's complexity, illustrating how different predictor combinations influence the model's behavior.

Demonstrating relationships of predictors and their impacts on model output

Visualization tasks. Participants articulated a preference for visualizing the relationships of predictors and their impacts on model outputs in CPM studies. A key visualization task includes *demonstrating how variations in one predictor* while keeping others constant, can affect model predictions (IDs 6, 8). This approach helps in understanding the isolated impact of each predictor on the model's risk predictions. Another important task is *showing how varying levels of continuous variables influence associated hazard or odds ratios* (ID 19). This visualization provides insights into the variable's impact across its range, which is crucial for comprehending complex relationships. Additionally, participants emphasized the need for visualizations that allow for the *manipulation of multiple predictors in realistic scenarios* (IDs 6, 13). Such visualizations enable users to adjust parameters like age and comorbidities simultaneously, offering a more accurate depiction of the model's performance in various real-life conditions. This multi-variable manipulation is essential for preventing misunderstandings that could arise from observing the effect of a single predictor in isolation. One participant aptly highlighted,

> "*I would want to know more about what is the interaction between X and Y? What is the interaction between Y and Z? Was Y standalone important clinically? Or is there something g*oing *on in the dynamics between the features that are clinically meaningful for me? ... It doesn't matter too much that I have a model that's 95% accurate versus 90% accurate, but if the 90% accurate model is able to touch upon these featured interactions and other clinical questions I have, I might find that a little bit more useful*." (ID 17)

Visual encoding. To visually depict the tasks mentioned above, several graphical techniques can be employed to provide a nuanced comprehension of predictor relationships and their impacts on model outputs. The *Tornado Diagram* (Briggs et al., 2012), for instance, neatly

organizes parameters along the Y-axis while illustrating their influence on model output along the X-axis, offering a sorted view based on the level of influence and enabling real-time 'what-if' analyses through interactive features. On the other hand, *Restricted Cubic Splines* provide a smooth representation of the relationships between variables and hazard/odds ratios, which is especially beneficial when the relationships are non-linear, thereby preserving the continuous nature of variables without the need for dichotomizing. Engaging with an Interactive *Parallel Coordinate Plot* could offer an intuitive understanding of how altering parameters like age and comorbidities influence the model's output, as users can adjust these parameters and observe the real-time changes in model predictions (Steed et al., 2012). The plot's vertical lines represent parameters, while polylines connect these axes, reflecting specific instances of parameter values. Lastly, the utilization of an *Interaction Effect Plot* with variations unveils the standalone importance of two-way interactions and facilitates model accuracy comparison by plotting predicted against actual outcomes, with color differentiation for distinct models (Lüdecke, 2023). Through these visualizations, the intricate dynamics of predictor relationships and their substantive impacts on model outputs can be explored comprehensively.

Assessing the models at different cut-offs or thresholds

Visualization tasks. Participants expressed a preference for visualization in presenting results of CPM studies, focusing on *assessing the models at different cut-offs or thresholds* to support clinical decisions (IDs 7, 11, 20). Visualization tasks in this context might include comparisons that illustrate the impact of applying specific probability cut-offs from model outputs, showing how patients would be categorized as having a disease or not. Clinicians can then assess the trade-offs of using such cut-offs, considering the potential to treat a reduced number of patients

and the corresponding changes in clinical practice (IDs 7, 11, 20). One participant encapsulated the essence of this analysis, stating,

"*So if you say, this is what the physicians are doing now. And if they use this model at a particular cut-off, this is what would have happened...*" (ID 20)

Visual encodings. A significant visualization encoding aiding this exploration is the *Decision curve plot for net benefit,* which contrasts the prevailing physician practice with the hypothetical employment of a prediction model at particular probability thresholds (Vickers et al., 2019). The x-axis represents the probability threshold or cut-off, while the y-axis illustrates the net benefit, thereby visualizing the relationships between different thresholds, the number of patients treated, and the net benefit. This graphical representation serves as a pragmatic tool for understanding the operational dynamics of CPMs across a spectrum of probability thresholds, aiding in the informed evaluation of model utility in real-world clinical scenarios.

6.4.2.5 Model performance

Model Performance encompasses a detailed and rigorous evaluation of the developed model beyond mere quantitative metrics. It is about assessing robustness and generalizability to ensure that the model is fit for its intended application. My study identified three visualization tasks preferred by participants in this section: *reporting model performance using multiple metrics, presenting more than just calibration plots, and comparing model performance.* Table 6.5 lists these tasks, along with their sub-tasks and visual encodings.

Table 6.5 Identified visualization for the model performance section

Visualization tasks - subtasks	**Visual encoding**	**#Visual encoding**
Reporting model performance using multiple metrics. (IDs 1, 6, 10, 11, 12, 13, 18, 19, 20)		

Visualization tasks - subtasks	**Visual encoding**	**#Visual encoding**
Evaluating discrimination ability	Receiver operating characteristic (ROC) curve	5-1
Assessing calibration	Calibration plot	5-2
Analyzing trade-offs	Decision curve analysis	5-3
Measuring general performance	Learning curves plot	5-4
Presenting more than just Calibration Plots (IDs 3, 7, 14, 18, 19)		
Assessing recalibration needs	Before-and-after comparison of calibration plots	5-5
Evaluating accuracy across risk levels	Detailed calibration plots with regions of interest highlighted	5-6
Comparing predicted vs. observed risk	Bar charts to compare predicted and observed risk	5-7
Comparing model performance (IDs 6, 18, 19)	Line graph to compare model performances	5-8

Reporting model performance using multiple metrics

Visualization tasks. Participants favored the inclusion of visualization in the presentation of results within CPM studies to allow for a comprehensive reporting of model performance across multiple metrics. The discussion around *evaluating discrimination ability* highlighted the importance of measuring the model's capacity to differentiate between distinct classes (IDs 3, 6, 11, 12, 13, 14). Concurrently, the emphasis on *assessing calibration* underscored the necessity to measure the agreement between predicted probabilities and actual outcomes, ensuring that the model's predictions are reliable and aligned with real-world data (IDs 7, 19, 20). Moreover, the discourse on *analyzing trade-offs* pointed towards evaluating and comparing prediction models by considering the balance between benefits (true positives) and harms (false positives), enriching the understanding of model performance in varied clinical scenarios (IDs 1, 2). Lastly, the broader theme of *measuring general performance* encapsulated the aspiration for a thorough assessment

of the model's predictive accuracy and adaptability to new data, fostering a deeper understanding of the model's utility in clinical practice (IDs 6, 18, 20). One participant encapsulated the essence of this comprehensive performance reporting by stating,

> "*But the (model) performance I mean the calibration not necessarily the discrimination, because I think when I look at just the ROC curve doesn't, doesn't mean that much. It can mean that I think the value of the AUC itself is enough to my, in my view*." (ID 7)

Visual encoding. To effectively illustrate the tasks mentioned, various visualizations can be utilized, each uniquely contributing to the understanding of the model's performance. The *Receiver Operating Characteristic (ROC)* curve is a crucial tool for assessing a model's discrimination ability. It plots the true positive rate against the false positive rate, with the Area Under the Curve (AUC) serving as a key metric for evaluation. The *Calibration plot* is another important graphical tool that compares the predicted probabilities of outcomes with the actual occurrences, thereby evaluating the accuracy of the model's predictions. Additionally, *Decision curve analysis* provides a graphical means to quantify the net benefits at varying threshold probabilities, offering a clear view of the trade-offs involved in sensitivity analysis (Vickers et al., 2019; Vickers and Holland, 2021). Lastly, *learning curve plot* also plays a vital role in visualizing a model's performance over both training and validation datasets as a function of the number of observations or iterations (Alharbey et al., 2022). This plot is particularly insightful in revealing how a model's learning capability and adaptability evolve with the introduction of more data, offering a comprehensive view of its learning trajectory.

Presenting more than just calibration plots

Visualization tasks. Participants expressed a preference for a more comprehensive presentation of model performances in reporting CPM study results beyond just showcasing

calibration plots. The dialogue around *assessing recalibration needs* accentuated the importance of evaluating whether models require recalibration, thereby pushing for a deeper exploration into model improvements (ID 3). The task of *evaluating accuracy across different risk levels* emerged as crucial, suggesting that merely presenting a calibration plot may not suffice in revealing discrepancies between predicted and observed probabilities across various risk strata (IDs 7, 14). Furthermore, the interviews highlighted the value of *analyzing risk-specific model performance*, urging a thorough investigation into the model's predictive accuracy across distinct risk ranges such as low, middle, and high (ID 14). The conversation also tilted towards *comparing predicted versus observed risk*, emphasizing that a mere calibration plot might overlook potential overestimation or underestimation of risk by the model, thereby advocating for a broader scope in the visual representation of model performance (IDs 18, 19). One participant highlighted the significance of this broader perspective, stating,

> "*But that rigorous testing is based on the calibration and the very few studies in the literature actually assess the calibration very well. They just plot the calibration and they said that's done. But in fact, assessing those calibrations and the results and to understanding like how this varies and do you need some kind of recalibration of the model depending on the that side or depending on the outcome in a different side."* (ID 3)

Visual encoding. To visually illustrate the tasks outlined above, a range of specific visualizations can be employed, each serving to enrich the comprehensive presentation of the calibration plots. The *before-and after comparison of calibration plots* acts as a clear visual indicator of any improvements achieved through recalibration or highlights areas where miscalibration persists, thereby shedding light on the model's evolving performance (Van Calster et al., 2019). *Detailed calibration plots with regions of interest highlighted* delve deeper by

segmenting the calibration plot into distinct risk zones, making the evaluation of discrepancies between predicted and observed probabilities more intuitive (Yin et al., 2022). The color-coding within these plots flags key areas, like low or high risk, where calibration may be off, offering a more nuanced understanding of the model's performance across different risk spectrums. Lastly, simple yet effective visualizations like *Bar charts* facilitate a direct comparison of predicted versus observed risk, enabling a straightforward assessment of the model's predictive accuracy (Norrish et al., 2023). By plotting each observed risk instance against its predicted risk, these plots help identify overlaps or discrepancies and assist in the identification of potential biases or systematic errors in the model.

Comparing model performance

Visualization tasks. Participants expressed a preference for CPM studies to go beyond presenting their own proposed models. The emphasis is on *comparing model performance*, particularly contrasting newly proposed models with existing models, baseline models, or standard practice. This includes quantifying and comparing models through visual representations, such as graphical devices that demonstrate true positives, false positives, true negatives, and false negatives, taking into account different weightings between false negatives and false positives between comparisons (ID 19). They also highlight the importance of visual comparisons, such as graphs, to show improvements in predictions between two models and compare the new model with the baseline or standard practice to demonstrate its potential advantages or simplifications (IDs 6, 18, 19). The need for clear comparisons extends to practical applications, as it affects the decision-making process when considering the adoption of one model over another in clinical practice. As one participant aptly put it,

"*Other graphical devices, things like the net benefit, which can take into account the weighting, different weighting between false negatives and false positives. I don't see that very often, but that can help explain where something might be useful more than, say, standard practice or other things.*" (ID 19)

Visual encoding. The utilization of visual encodings for comparing model performances may convey complex insights. *Line Graph Utilization* serves this purpose by employing different line styles to indicate each model and its corresponding weighted accuracies (Pennello et al., 2016). The graph is structured with the y-axis representing accuracy and the x-axis representing the cost/benefit ratio 'r.' Including a separate line that portrays the difference in performance between the models further emphasizes the comparative aspect, and the addition of a legend ensures user-friendly interpretation.

6.4.2.6 Model presentation

As CPM studies conclude the process of developing a CPM, the report of the result arrives at the pivotal stage of model presentation. This final stage is where the model, in all its complexity, is distilled into a clear and comprehensible format (Steyerberg and Vergouwe, 2014). This stage serves as the presentation of the model for target users to potentially apply in their contexts. Table 6.6 provides a summary of visualization tasks, their sub-tasks, and visual encodings for this section.

Table 6.6 Identified visualization tasks for the model presentation

Visualization tasks - subtasks	Visual encoding	#Visual encoding
Presenting simplified model presentation (IDs 11, 18, 19, 6)		
Translating complex models into scoring systems	Nomogram	6-1

Visualization tasks - subtasks	Visual encoding	#Visual encoding
Visualizing diagnostic values that integrate pre-test probabilities, likelihood ratios obtained from the prediction models, and post-test probabilities.	Fagan's diagram	6-2
Utilizing graphical symbols to represent statistical probabilities	Icon arrays or Pictorial Units (like human figures)	6-3
Optimizing model presentation through interactive visualization that engage the users (IDs 6, 7, 9, 10, 13, 14, 18, 19, 20)	Various interactive visual applications such as web-based calculator	6-4

Presenting simplified model presentation

Visualization tasks. Participants preferred simplified presentation of results in CPM papers, aiming for a more intuitive understanding and user-friendly application. Participants specifically showed a tendency towards *translating complex models into scoring systems*, facilitating a more straightforward interpretation and application of the models in clinical settings (IDs 6, 18). Moreover, the conversation highlights *visualizing diagnostic values that integrate pre-test probabilities, likelihood ratios from the models, and post-test probabilities*, enriching the narrative on the model's diagnostic utility (ID 11). Participants also mentioned *utilizing graphical symbols to represent statistical probabilities*, making the data more tangible and interpretable to both clinicians and patients (ID 19). These simplified model presentations can bridge the gap between complex statistical outcomes and practical clinical utility. As one participant emphasized, translating complex models into simpler formats is essential as

"*Well, again, so maybe that Fagan's diagram or showing how, for example, I don't know if a certain disease requires treatment above, like putting that in perspective with the next steps again, right? Whether it is a diagnostic test and then it could be a Fagan's nomogram of how such a pre-test will be influenced by the result of the diagnostic test and what will be the post-test probability of disease or some discussion of risk benefit of treatment and above which probability*

of disease the chances is that patient will get more benefits at harm from being treated for that disease. " (ID 11)

Visual encoding. In visualizing the tasks mentioned above, visual tools can be employed to showcase the details of the presentations of CPMs. *Nomograms* can showcase the transformed complex mathematical models into straightforward, visual forms based on point-based scores for easier interpretation and practical application. On another front, *Fagan's diagram*, with its intersecting lines on linear scales, aids users in deriving post-test probabilities from pre-test probabilities and likelihood ratios, making the information more digestible. Similarly, icon arrays utilize *pictorial units like human figures* to represent probabilities visually, making statistical data more relatable and understandable. For instance, a 10% risk could be illustrated by coloring 10 out of 100 figures, providing a clear, visual representation of probabilities that can be easily grasped by both clinicians and patients. Through these visualization tools, the data from CPMs is not only made more accessible but also more engaging for its intended audience.

Optimizing model presentation through interactive visualization that engage the users

Visualization tasks. Participants highlighted the vital role of *interactive visualization tasks* in presenting results within CPM papers. The core concept hinges on the provision of interactive tools to enrich the model's presentation. These interactive mediums enable users to input data and receive real-time feedback, rendering the models more user-friendly and easier to grasp (IDs 7, 9). For example, the creation of online calculators or applications where users can input patient-specific information and receive personalized risk assessments was highly advocated as it enhances model understandability and utility (IDs 10, 14, 19). This method not only facilitates a deeper understanding of the model but also provides a hands-on experience, allowing users to discern the real-time implications of varying input values on the outcome (ID 20). Moreover,

participants indicated that the transition from paper to interactive digital platforms markedly enhances the model's usability. They encouraged the development of supplementary tools such as Excel sheets, Shiny apps, or web-based platforms for a more engaging and user-oriented model presentation, albeit with a caveat (ID 18). As one participant insightfully pointed out,

> "*I think it should be close to for the presentation of the model I think should be close to what the user would be provided with. Like, I like the idea of, and some of the papers already showed that. Like, web, web-based calculator, for example, online calculator. It is nice in a way that one can see the different items. One can see how to fill in the different items, and then one can see that is written something*." (ID 7)

Visual encodings. Interactive visual applications can be deployed to visualize the tasks mentioned previously, utilizing tools that take on user input. These applications could be presented as web applications or on other similar platforms, making them easily accessible to a wide range of users. This interactive dimension provides a tangible way for users to interact with the model, observe real-time modifications based on their input, and gain a better understanding of the model's behavior and output. By allowing users to alter variables and view the immediate impact on outcomes, these applications provide a user-centric approach and a practical implementation of the models in clinical decision-making. This step towards interactive visualization underscores the evolving narrative in presenting results of CPM studies, where user engagement is at the forefront.

6.4.2.7 Multi-sections

Confidence intervals play a pervasive role in the development of CPMs, including stages such as data exploration, feature selection, modeling, model performance, and result exploration. Rather than being specifically allocated to individual sections, confidence intervals have been

categorized under ‘multi sections.’ The visualization task for this section is only one about integrating the uncertainty measurements into visualizations, as depicted in Table 6.7.

Table 6.7 Identified visualization task for the multi-section

Visualization tasks - subtasks	Visual encoding	#Visual encoding
Representing the variability and uncertainty of measurements to aid in the interpretation of results. (IDs 13, 19)	Incorporating confidence intervals into various types of plots depending on the context	7-1

Integrating the uncertainty of measurements

Visualization tasks. Participants emphasized the importance of *representing the variability and uncertainty of measurements to aid in the interpretation of results* (IDs 13, 19). This visualization task should focus on making abstract data from CPM study results more concrete and interpretable by highlighting how measurement uncertainties can influence the understanding of the results. A key quote reflecting this sentiment is. A key quote reflecting this sentiment is:

> "*I think what's really important, and in fact is more important than the point estimates, is the, what I call measurement error, which is usually a confidence interval of some kind. And that's also shown, may be shown with a graph sometimes. But, it is that when I see a result, and sometimes I see, I often see this in an abstract, where the results are presented only the point estimates and not the confidence interval. The confidence interval is more important than the point estimate.*" (ID 19)

Visual encoding. Representing the variability and uncertainty of measurements through visualizations involves the integration of confidence intervals, which are centered on visually illustrating the uncertainty tied to various estimates (Han et al., 2009; Haskins et al., 2013). This approach provides viewers with a comprehensive understanding of the results, extending beyond mere point estimates to showcase the plausible range of these estimates and offering insights into

potential variations. Confidence intervals can be integrated into various types of plots depending on the context, such as adding error bars to bar graphs, ribbons around line graphs, or whiskers on box-and-whisker plots. This inclusion not only enriches the interpretation of the data but also enhances its trustworthiness by demonstrating the plausible range of outcomes rather than relying solely on single point estimates.

6.4.3 *RQ3: How can visualizations preferred by biomedical researchers ensure understandable, useful, and trustworthy results in CPM studies?*

During the interviews, I found that participants acknowledged the potential of various visualizations to enhance the quality of CPM study results across the themes of understandable, useful, and trustworthy, particularly when discussing how visualizations could augment these qualities. Table 6.8 presents a list of visualization tasks suggested by participants, along with their corresponding themes, sub-themes, and #Needs that I identified during the qualitative analysis of needs (Chapter 5). This table is based on Appendix I that provides the complete participant answers used for the analysis for this research question. #Needs refer to the identified needs among participants to ensure understandable, useful, and trustworthy results of CPM studies. The complete list of # Needs is reported in Chapter 5.

6.4.3.1 Visualizations to support understandable results

My analysis found that participants considered the use of visualizations (visualizations tasks and encodings) to support understandable results in CPM studies, specifically within three of the six sub-themes of 'understandable' theme: 'comprehension,' 'analysis,' ‘application’, and 'evaluation.' Table 6.8 shows the compilations of visualization tasks and their corresponding themes of ‘understandable’ and sub-themes.

Table 6.8 Visualization tasks and their corresponding themes (‘understandable’), sub-themes, and # Needs

Themes & Sub-themes	Needs	Section	Visualization Tasks	Visual Encoding
Understandable - Comprehension	Ensuring user-centric reporting and calling for the study results to be understandable to a broad audience, including those with limited technical expertise, especially clinicians as the target users. (#1-2-1)	Model performance	Reporting model performance using multiple metrics	- Receiver operating characteristic (ROC) curve, - Calibration plot, - Decision curve analysis, - Learning curves plot
	Making the analysis using machine learning models interpretable that avoid the "black box" problem, and provide clear, comprehensive explanations of complex methods. (#1-2-2)	Modeling	Explaining model architecture	- Decision trees or rule sets - Model diagrams
Understandable – Analysis	Breaking down data exploratory phase that delves into the data discovery phase, showcasing initial data signals, and detailing specific data characteristics that could influence model building and interpretation. (#1-3-1)	Data exploration	Detecting initial data patterns, trends, or anomalies, outliers	- Scatter plots - Histogram
	Showcasing the complex model relationships between variables and how they contribute to the overall model output and performances. (#1-3-4)			

Themes & Sub-themes	Needs	Section	Visualization Tasks	Visual Encoding
	Describing the detailed steps of model development. (#1-3-2)	Predictor selection	Visualizing predictor selection	- Correlation matrix heatmaps - Tables with predictors, coefficients, and intercepts - Flowcharts or process diagrams
	Showcasing the complex model relationships between variables and how they contribute to the overall model output and performances. (#1-3-4)	Result exploration	Highlighting predictor importance,	- Ordered list or bar chart
			Demonstrating relationships of predictors and their impacts on model output	- A forest plot - Interaction plots
	Analyzing model performance using multiple metrics. (#1-3-3)	Model performance	Reporting model performance using multiple metrics	- Receiver operating characteristic (ROC) curve, - Calibration plot, - Decision curve analysis, - Learning curves plot

Themes & Sub-themes	Needs	Section	Visualization Tasks	Visual Encoding
			Presenting more than just calibration plots	- Before-and-after comparison of calibration plots - Detailed calibration plots with regions of interest highlighted - Bar charts to compare predicted and observed risk
Understandable - Application	Allowing users to try out prediction models through clear operational guidance, user-friendly scoring systems, and the use of web-based applications for active testing and implementation in clinical settings. (#1-4-1)	Model presentation	Presenting simplified model presentation	- Nomogram *Fagan's diagram*
			Optimizing model presentation through interactive visualization	- Fagan's diagram - Icon arrays or Pictorial Units (like human figures) - Various interactive visual applications

Themes & Sub-themes	Needs	Section	Visualization Tasks	Visual Encoding
Understandable - Evaluation	Comparing the performance of proposed models with the other existing models or standard practices. (#1-6-3)	Result exploration	Assessing models at different cut-offs or thresholds	- Decision curve plot for net benefit
	Comparing the performance of proposed models with the other existing models or standard practices. (#1-6-3)	Model performance	Comparing model performances	- Line graph to compare model performances

Understandable - Comprehension

For comprehension, participants emphasized the need to consider the users when presenting CPM study results. For example, when discussing the visualization tasks of reporting model performance using multiple metrics, participants highlighted the importance of tailoring visual data presentations to the background and needs of the target audience, ensuring that the results are both interpretable and relevant to the intended users (ID 18). Participants suggested using visual encodings such as ROC curves and calibration lines. While ROC curves are popular, they can be challenging for clinicians unless clearly marked. In contrast, calibration lines are easier to understand visually than text. Participants also recommended providing context through text beyond figure legends and table captions to prevent misinterpretation (ID 12). One participant illustrated this by saying,

> "*I know people like ROC curves and things like that. I think the average clinicians, they find that those quite tough unless they're very clearly marked. So, yeah, whereas calibration, that the lines are far easier to understand than the text. So, as an example, I'd say there are some things that are very clearly visual, like a calibration line, and some things that need a good example, very clear non-technical example. And I'd say a discrimination statistic is in need of that.*" (ID 12)

Another area where participants advocated for better contextual information is in presenting study results from CPM studies that use machine learning models. Participants emphasized the importance of making the analysis using machine learning models interpretable to avoid the "black box" problem and provide clear, comprehensive explanations of complex methods (IDs 12, 17). Participants further pointed out that while it may be challenging to showcase how

machine learning models differ from simpler representations, it is crucial to strive for transparency in explaining the components of the methodology and how variables are selected (IDs 6, 12). To address these issues, visual encodings such as decision trees or rule sets can be used to illustrate how the model makes predictions, as these can provide a step-by-step breakdown of the decision-making process during the modeling stage of developing CPMs using machine learning models. Model diagrams can also represent the structure and relationships within the model, helping to clarify the modeling stages. One participant summarized this need by stating,

> "*But maybe more of an issue around explaining components of the methodology and how variables were selected or in particular and increasingly these days what goes in the black box of machine learning or artificial intelligence type tools, which is the particular problem, ... That's where I would want quite explicit criteria of either understandability or the researchers owning up and being clear that actually there's a point if you're doing a black box type of neural network or so on that that might not be directly understandable and we have to take it on faith.*" (ID 14)

Understandable - Analysis

For ‘analysis’, which focuses on participants' desire for CPM study results to break down concepts and understand their relationships, participants expressed a need for detailed data exploration. Participants further emphasized that in this stage, authors should showcase initial data signals and specific characteristics that could influence model building and interpretation (ID 17). Additional emphasis was on the importance of illustrating complex relationships between variables and their impact on model output and performance (ID 14). To address these needs, suggested visualization tasks may include scatter plots for detecting non-linearity, boxplots or violin plots

for outlier detection, histograms for feature characteristics, and various plots to show normality distributions before and after transformation. One participant stated,

> "*You have to be able to explain that other things going on behind the model of variable that seems like it shouldn't behave the way it is makes it behave that way in terms of, you know, how much variability there was in it relative to other other variables in the model or whatever it is, you know, it is even hard for the statistician.*" (ID 14)

The other stage of presenting CPM study results that participants suggested breaking down into parts to understand their relationships is predictor selection. This approach fulfills the participants' need to describe the detailed steps of model development. Suggested visualizations for this stage include correlation matrix heatmaps, tables with predictors, coefficients, and intercepts, and flowcharts or process diagrams. These visualizations help clarify how predictors were chosen and their relevance to the model (ID 18). For instance, one participant emphasized the importance of transparency in this stage:

> "*And it shouldn't be like, you shouldn't go into too much trouble to have a fancy figure for that. It may be as simple as a table with all coefficients. And also, sometimes it is not even clear, like, what the final predictors were, because many, many times they start with, I don't know, 100 predictors and then they skip those that were significant in the deep array dissociation and so on and so on.*" (ID 18)

Result exploration is also the stage where participants suggested breaking down concepts into parts and understanding their relationships. In this stage, it is crucial to showcase the complex model relationships between variables and how they contribute to the overall model output and performance (ID 13). To do this effectively, authors can highlight predictor importance and demonstrate the relationships of predictors and their impacts on the model output in presenting

results of CPM studies. Suggested visualizations for this stage include ordered lists or bar charts to show the ranking of predictor importance, forest plots to display the effects of predictors, and interaction plots to illustrate how different predictors interact and influence the model. One participant highlighted this need by stating,

> "*And another option is if it is not that primary goal to have an exact prediction model but have a model to say which factor has which importance. So for example, you have like 10 different predictors and you want to see what is the ranking, what is the most important one, and so on and so on.*" (ID 13)

Another stage where participants suggested breaking down concepts into parts to understand their relationships is model performance. In this stage, analyzing model performance using multiple metrics is crucial (IDs 3, 7, 13, 20). To achieve this, reporting model performance can use various metrics accompanied by their visual encodings such as the receiver operating characteristic (ROC) curve, calibration plot, decision curve analysis, and learning curves plot. Furthermore, participants noted that while the area under the curve (AUC) is commonly reported, it is equally important to include variations of calibration plots to provide a comprehensive performance analysis (IDs 7, 14, 19). These variations may include visual encodings such as detailed calibration plots with regions of interest highlighted, risk-stratified performance plots, and bar charts to compare predicted and observed risks. For example, one participant emphasized the necessity of these visualizations by stating,

> "*Well, I mean, I think that, you know, having calibration plots are helpful, you know, and the AUC curve is helpful for people that know what it is, but it is not helpful for people who don't know what it is. And yeah, often you don't see anything about calibration, which is obviously a very important piece of it.*" (ID 20)

Understandable - Application

For the ‘application’ sub-theme, participants discussed the need for results to demonstrate the application of CPMs in clinical settings. To achieve this, in the model presentation stage, authors can allow users to try out prediction models through clear operational guidance, user-friendly scoring systems, and the use of web-based applications for active testing and implementation in clinical settings (IDs 6, 11). These can be achieved by presenting simplified model presentations and optimizing them through interactive visualization. Suggested visualizations include nomograms, Fagan's diagrams, icon arrays or pictorial units (like human figures), and various interactive visual applications. One participant emphasized the importance of simplicity and clarity in model presentation by stating,

> "*Is that sufficient to be understandable, or does it have to be a really simple rule, like, you know, a simple decision list or some sort of scoring system where it is really easy, or it is understandable both what goes into it and how the model is operating*." (ID 6)

Understandable - Evaluation

For ‘evaluation’, participants want the results to help them judge the value of presented information backed by evidence. Participants suggested that in the result exploration stage, authors can compare the performance of proposed models with other existing models or standard practices (IDs 19, 6). This comparison can be done by assessing models at different cut-offs or thresholds and using visualizations such as decision curve plots for net benefit. This visual encoding can help explain where a model might be more useful than standard practices, as they take into account different weightings between false negatives and false positives. One participant noted,

"*The other graphical devices, things like the net benefit, can take into account the weighting between false negatives and false positives. I don't see that very often, but it can help explain where something might be useful more than, say, standard practice.*" (ID 19)

6.4.3.2 Visualizations to support useful results

The study participants suggested various visualization tasks and encodings within the context of a 'useful' theme, particularly under the sub-themes of 'critical support,' 'easy to use,' 'useful for tasks,' and 'user-control.' Table 6.9 shows the compilations of visualization tasks and their corresponding themes of ‘useful’ and sub-themes.

Table 6.9 Visualization tasks and their corresponding themes (‘useful’), sub-themes, and # Needs

Themes & Sub-themes	Needs	Section	Visualization Tasks	Visual Encoding
Useful - Critical support	Guiding whether or not to use the prediction models by appropriately clarifying their intended function and outlining potential limitations and risks to avoid misuse. (#2-1-1)	Result exploration	Highlighting predictor importance,	- Ordered list or bar chart - A forest plot - Interaction plots
Useful - Easy to use	Convincing users that the prediction models are easy to use by portraying prediction models as straightforward and user-friendly, typically through intuitive tools like nomograms and interactive digital platforms. (#2-2-1)	Model presentation	Presenting simplified model presentation, Optimizing model presentation through interactive visualization	- Nomogram - Fagan's diagram - Icon arrays or Pictorial Units (like human figures) - Various interactive visual applications
Useful - Useful for tasks	Demonstrating the models’ robustness and how the models work. (#2-5-1)	Result exploration	Highlighting predictor importance	- Ordered list or bar chart - A forest plot - Interaction plots
			Demonstrating relationships of predictors and their impacts on model output	- Tornado diagram - Restricted cubic splines Interactive parallel coordinate plot - Interaction effect plot with variations - Plotting standalone importance

Themes & Sub-themes	Needs	Section	Visualization Tasks	Visual Encoding
				- Plotting two-way interactions Comparing model accuracy
			Assessing models at different cut-offs or thresholds	- Decision curve plot for net benefit.
	Demonstrating the models' robustness and how the models work. (#2-5-1)	Model performance	Presenting more than just calibration plots	- Before-and-after comparison of calibration plots - Detailed calibration plots with regions of interest highlighted - Bar charts to compare predicted and observed risk
Useful - User-control	Empowering readers to have more control over their required or desired tasks, suggesting the incorporation of interactive simulations and hypothetical scenario analyses for enhanced engagement with the study findings. (#2-8-1)	Result exploration	Highlighting predictor importance	- Ordered list or bar chart - A forest plot - Interaction plots
			Demonstrating relationships of predictors and their impacts on model output	- Tornado diagram - Restricted cubic splines Interactive parallel coordinate plot - Interaction effect plot with variations

Themes & Sub-themes	Needs	Section	Visualization Tasks	Visual Encoding
				- Plotting standalone importance - Plotting two-way interactions Comparing model accuracy

Useful – Critical support

. For ‘critical support’, participants expect the results to present how the prediction models would support critical elements of the end users' required or desired tasks; otherwise, the end users or readers may fail at their tasks if the results don't deliver such presentations. The tasks that authors can make clear for readers to prevent failure include highlighting or clarifying potential limitations and risks of the models, and emphasizing predictor importance, particularly during the result exploration stage (ID 6). To convey these presentations, recommended visual encodings include ordered lists or bar charts for ranking predictors, forest plots for comparing predictor effects, and interaction plots to show how predictors interact and whether they are harmful or protective. One participant noted,

> "*I think having a top five or top three, for both harmful and protective effects, or just top five by magnitude and indicating whether they're harmful or protective, is a good way to present results*." (ID 6)

Useful – Easy to use

For 'easy to use', participants want the results to show that it will be easier for end users to use CPMs when performing their tasks or work. To address this need, participants expressed their preference for authors to convince users or readers that the prediction models are straightforward and user-friendly, typically through intuitive tools (IDs 18, 19). This is especially important in model presentations, where simplified model presentations and interactive visualizations can make a significant difference. Suggested visualizations include nomograms, Fagan's diagrams, icon arrays or pictorial units (like human figures), and various interactive visual applications. One participant emphasized the importance of usability by stating,

"*If that is your target population, I would use a risk chart or a nomogram or translate your model into a point-based model. If you have this at one, you have that at two, and so on. It is not ideal from the statistical point of view, but at least it is going to be usable or most likely to be usable in those contexts*." (ID 18)

Useful - Useful for tasks

For 'useful for task', participants want the results to be useful for readers' tasks as researchers or clinicians. As readers, it is important to see demonstrations of the models' robustness and how the models work (IDs 6, 13). This can be achieved in the results exploration stage by highlighting predictor importance using visualizations such as ordered lists or bar charts, forest plots, and interaction plots. Additionally, demonstrating relationships of predictors and their impacts on model output is crucial (IDs 13, 17), which can be visualized through tornado diagrams, restricted cubic splines, interactive parallel coordinate plots, interaction effect plots with variations, plotting standalone importance, plotting two-way interactions, and comparing model accuracy. Assessing models at different cut-offs or thresholds using visual encodings like decision curve plots for net benefit also adds value (ID 17). In summary, participants noted that such detailed visual encodings make the models more practical and relevant for their tasks as readers. One participant emphasized this by stating,

" *So, for me, it is very unsatisfactory not to know how the work in between goes. So, the models that we create, we list the influence factors and we show the effect of each factor*. " (ID 1)

Regarding the models' robustness, what readers need to see can be reflected through model performance. Participants emphasized the importance of presenting more than just calibration plots to fully demonstrate model performance (IDs 3, 18). Effective visualizations include before-and-

after comparisons of calibration plots, detailed calibration plots with regions of interest highlighted, risk-stratified performance plots, and bar charts to compare predicted and observed risk. These visualizations help readers understand the nuances of model calibration and its variations across different settings, ensuring that the models are accurately evaluated, and their performance is clearly communicated to the readers. One participant highlighted the necessity of rigorous calibration assessment by stating,

> "*But that rigorous testing is based on the calibration and the very few studies in the literature actually assess the calibration very well. They just plot the calibration and they said that's done. But in fact, assessing those calibrations and the results and to understanding like how this varies and do you need some kind of recalibration of the model depending on the that side or depending on the outcome in a different side."* (ID 3)

Useful - User-control

For 'user control', participants want to empower readers to have more control over their tasks, suggesting the incorporation of interactive simulations and hypothetical scenario analyses for enhanced engagement with the study findings. Participants expressed the importance of allowing readers to interact with the data, especially in the results exploration stage, by highlighting predictor importance using visualizations such as ordered lists or bar charts, forest plots, and interaction plots (IDs 6, 13). Additionally, demonstrating relationships of predictors and their impacts on model output can be achieved using visual encodings such as tornado diagrams, restricted cubic splines, interactive parallel coordinate plots, interaction effect plots with variations, plotting standalone importance, and plotting two-way interactions to compare model accuracy (IDs 6, 2). One participant emphasized this by stating,

"*Ideally, with modeling, you would want to see some kind of almost even a simulation where the expert then explains if you change some of the entry parameters or model assumptions, then you get results like this. This means that this factor is important, other ones are not.*" (ID 2)

6.4.3.3 Visualizations to support trustworthy results

Participants identified visualization tasks and encodings related to the theme of 'trustworthy,' with a particular focus on the sub-themes of 'accuracy,' 'objectivity,' and 'validity.' Table 6.10 shows the compilations of visualization tasks and their corresponding themes of ‘trustworthy’ and sub-themes.

Table 6.10 Visualization tasks and their corresponding themes ('trustworthy'), sub-themes, and # Needs

Themes & Sub-themes	Needs	Section	Visualization Tasks	Visual Encoding
Trustworthy - Accuracy	Ensuring accuracy in data presentation that avoids mistakes such as discrepancies in displayed results and misleading visual representations. (#3-1-1)	Data exploration	Detecting initial data patterns, trends, or anomalies, outliers	- Scatter plots - Histogram
Trustworthy - Objective	Justifying the decisions made during the research process, which includes the careful selection of analytical methods and variables, confronting and addressing anomalies like outliers, detailed methodology disclosure, handling missing data, determining sample size, and avoiding overly aggressive methods. (#3-2-3)	Data exploration	Detecting initial data patterns, trends, or anomalies, outliers	- Scatter plots - Histogram
	Justifying the decisions made during the research process, which includes the careful selection of analytical methods and variables, confronting and addressing anomalies like outliers, detailed methodology disclosure, handling missing data, determining sample size, and avoiding overly aggressive methods. (#3-2-3)	Predictor selection	Visualizing predictor selection	- Correlation matrix heatmaps - Tables with predictors, coefficients, and intercepts - Flowcharts or process diagrams
	Substantiating the model performance and improvement claimed using multi-dimensional assessment methods and benchmarking against existing models. (#3-2-4)	Model performance	Reporting model performance using multiple metrics	- Receiver operating characteristic (ROC) curve, - Calibration plot, - Decision curve analysis,

Themes & Sub-themes	Needs	Section	Visualization Tasks	Visual Encoding
				- Learning curves plot
			Comparing model performances	- Line graph to compare model performances
	Acknowledging limitations inherent in study results that include acknowledging the importance of confidence intervals over point estimates, addressing potential biases, and underscoring the role of calibration, alongside discrimination. (#3-2-1)	Multi-section		- Incorporating confidence intervals into various types of plots depending on the context
Trustworthy - Validity	Applying proper validation to the models. (#3-3-2)	Model performance	Reporting model performance using multiple metrics	- Receiver operating characteristic (ROC) curve, - Calibration plot, - Decision curve analysis, - Learning curves plot

Trustworthy - Accuracy

For accuracy, participants highlighted the importance of the information in CPM papers being error-free and supported by appropriate data. Ensuring accuracy in data presentation is crucial to avoid mistakes such as discrepancies in displayed results and misleading visual representations (ID 19). Specifically, participants emphasized the need to ensure accuracy when detecting initial data patterns, trends, anomalies, and outliers during data exploration which can be achieved through the accurate use of visual encodings such as scatter plots and histograms in the data exploration. By accurately presenting data distributions and variations, researchers can avoid misinterpretations and provide a clearer understanding of the data. One participant illustrated this need by stating,

> "*In table one, where you're just describing the demographics, or other information, if they present, say a blood measurement, and they present a mean of 100, and a standard deviation of 300, I don't trust that, because I know it is not normally distributed from that. If you take one or two standard deviations from the mean, you get a negative concentration of your blood, which is nonsense. So, they don't really understand, you need to present a median, and interquartile range instead, or maximum, minimum as well, or do it graphically, so you can see the distribution.*" (ID 19)

Trustworthy - Objective

For objectivity, participants advocated for results to be free from bias, deception, and distortion, ensuring a balanced view. They stressed justifying research decisions, especially during data exploration where anomalies like outliers are addressed and appropriate analytical methods are selected (IDs 3, 17). During this stage, authors can present initial patterns, trends, and outliers

detected using scatter plots and histograms, helping justify their decisions and move to the next steps. Objective visualizations can help represent unusual data points and ensure transparency of the follow-up justification. One participant noted,

> "*It is very important to show all the groundwork that we do in terms of data cleaning, data pre-processing, and checking the data for all the assumptions and then fixing... it is difficult to author this because... most of this just gets summarized in one single sentence or two sentences at best.*" (ID 17)

Ensuring the study results are objective by justifying decisions made is also emphasized during predictor selection. Participants expressed the importance of authors clarifying the rationale behind the selection of predictors and ensuring that the chosen predictors are independent and relevant to the model (ID 17). To address this, authors can use visualizations such as correlation matrix heatmaps, tables with predictors, coefficients, and intercepts, and flowcharts or process diagrams to show their justification for selecting predictors. Unfortunately, demonstrating objectivity is often overlooked at this predictor selection stage. As one participant stated,

> " *I find most clinical papers do not go into issues with multicollinearity, which is a very important aspect about linear regression. You want to have all your features being independent. I feel like as a field, we don't go beyond to try to come up with data visualizations that show more properties of the data that helped us build those predictive models.*" (ID 17)

Participants also indicated that presenting the study results as objective can be achieved by substantiating the model performance and improvement claimed using multi-dimensional assessment methods and benchmarking against existing models (IDs 6, 18). To address this, participants suggested reporting model performance using multiple metrics, including visual

encodings such as the receiver operating characteristic (ROC) curve, calibration plot, decision curve analysis, and learning curves plot. Additionally, comparing model performances using line graphs can effectively highlight the differences and improvements over existing models. This thorough approach underlines the necessity of detailed and transparent performance reporting to demonstrate the robustness and superiority of new models objectively, which also supports informed decision-making among users. One participant stated,

> "*If you're reading a risk prediction model paper and you are thinking of using that in your patients or in your practice or for something what pieces of information do you look for to make your decision of using model A or model B? And once you answer that for yourself we should try to always report that.*" (ID 18)

Presenting study results by acknowledging limitations inherent in the models includes recognizing the importance of confidence intervals over point estimates, which could address potential biases. This can help convince users that the study results are objective. To achieve this, incorporating confidence intervals into various types of plots depending on the context is essential (IDs 13, 19). For example, confidence intervals can be added to ROC curves, odds ratio plots, and other visual encodings to enhance the understanding of predictor importance and model reliability. By presenting confidence intervals, authors provide a clearer picture of the variability and reliability of their results, helping to mitigate potential misinterpretations based on point estimates alone (ID 19). One participant highlighted this by saying,

> "*But, the trustworthiness. How can it be produced? I think what's really important, and in fact is more important than the point estimates, is the, what I call measurement error, which is usually a confidence interval of some kind.*" (ID 19)

Trustworthy - Validity

For validity, participants discussed the importance of using accepted practices in reporting CPM study results, emphasizing proper model validation with multiple metrics (ID 20). Reporting model performance should include visual encodings like ROC curves, calibration plots, decision curve analysis, and learning curves. They noted that omitting calibration plots can lead to misunderstandings about a model's applicability to new datasets. Including these metrics ensures transparency in both discrimination and calibration, aiding users in understanding the model's reliability across different datasets. One participant highlighted,

> "*If you don't include a calibration plot, people may not understand that the model won't be calibrated for their own data... the discrimination might be the same, but the calibration might be really different.*" (ID 20)

6.5 DISCUSSION

This study characterizes the preferences of biomedical researchers regarding the use of visualizations in presenting results of CPM studies that are understandable, useful, and trustworthy. Findings reveal a strong preference among participants for employing visualizations in the dissemination of CPM findings and accompanying those visualizations with text to give contexts (RQ1). In RQ2, participants demonstrated preferences for visualization tasks and encodings across six stages of CPM development and validation (preimpact analysis studies), plus a multi-section focusing on incorporating confidence intervals. Finally, in RQ3, the study underscores the role of these visualization tasks and encoding in enhancing the understandability, usefulness, and trustworthiness of CPM study outcomes.

Preferences for using visualizations in presenting results of CPM studies

The exploration of visualization techniques in the presentation of information has predominantly centered on decision-making contexts, particularly for clinicians and patients (Eberhard, 2023; Lor et al., 2019). However, its application in the domain of research papers, especially in conveying complex data effectively, remains less explored. This study breaks new ground by demonstrating the feasibility and importance of studying preferences for visualizations among biomedical researchers in the context of CPM research papers. A key takeaway from this study is the unanimous agreement among participants on the vital role of visualizations, including tables, figures, and other visual aids, in effectively communicating complex data in CPM studies. This finding aligns with the broader trend of incorporating visual elements in various information dissemination among academic research community (Ma et al., 2012).

However, this study also brings to light the crucial need for a balanced approach to the use of visualizations. Participants emphasized that visual elements should not stand alone or be used without careful consideration. The accompanying text is paramount in providing context and preventing misinterpretation of the visual data. This symbiotic relationship between text and visuals is essential for a comprehensive and accurate presentation of study results. Participants cautioned against the overuse or misinterpretation of visuals, noting that excessive complexity or misleading representations could confuse and mislead the audience. This warning highlights the importance of judicious use of visualizations, ensuring that they serve to clarify rather than obfuscate the research findings.

Mapping visualization preferences for six sections of presenting results in CPM studies

This study builds upon previous research, which has outlined various visualizations used in predictive models and categorized them into six distinct sections: data exploration, predictor

selection, modeling, result exploration, model performance, and model presentation (Lu, 2017; Steyerberg and Vergouwe, 2014). While earlier studies have identified these visualizations through literature reviews, this current research uniquely contributes by capturing and analyzing participants' preferences for visualization tasks within these same sections of CPM studies.

The key achievement of this study lies in its detailed mapping of participant preferences for specific visualization tasks and visual encodings within each of the six stages of preimpact analysis studies. Interestingly, most of the visualization tasks mentioned by participants could be translated into visual encodings, drawing upon existing literature. This suggests that participants have a grounded understanding of visualization use in presenting results of CPM studies, which they have effectively applied in expressing their preferences. This comprehensive identification process has not only confirmed the visualizations identified in previous literature but also enriched them with direct insights from participants. Thus, this study provides a more user-centric perspective on visualization in presenting results of CPM studies, bridging the gap between theoretical knowledge from literature and practical preferences from the field.

Visualizations to support understandable, useful, and trustworthy results

This study has identified and mapped a range of visualization tasks and encodings based on participant preferences that span across the three critical themes of 'understandable,' 'useful,' and 'trustworthy' in presenting results of CPM studies. The findings demonstrate that each theme necessitates specific types of visualizations to meet its unique requirements, highlighting the multifaceted nature of conveying complex information to ensure understandability, usefulness, and trustworthiness in CPM studies.

For 'understandable' results, the emphasis is on clarity and broad accessibility. This ensures that complex data and models are presented through visualizations that are easily grasped by a

wide range of audiences, including those with limited technical expertise. This aspect is particularly crucial in the context of CPMs, where the ability to convey intricate details in an accessible manner can significantly impact the understanding and application of these models in clinical settings.

For 'useful' results, the focus shifts to practicality and user engagement. Visualization should enhance the practical application and operational integration of CPMs into clinical workflows. This includes interactive and user-friendly visualizations that not only explain the models but also facilitate their application in real-world settings.

In contrast, ensuring 'trustworthy' results through visualizations demands a keen emphasis on accuracy and objectivity. The study identifies visualization tasks and encodings that support the credible presentation of data and findings, underpinning the reliability and validity of CPM studies. These visualization strategies are essential for building trust and confidence in the models among clinicians, researchers, and decision-makers.

Future Studies

The present study's findings on visualizations in CPM papers open new pathways for further investigation. Specifically, future studies could explore the integration of Human-Centered Design (HCD) concepts in developing and presenting visualizations to report CPM study results (FDIs, 2009). The participants' emphasis on making study results more understandable, useful, and trustworthy points toward the potential of employing HCD principles. By prioritizing the end-users’ needs and preferences, HCD can guide the development of visualizations that are both intuitively grasped and scientifically robust. This approach could also extend to the examination of interactivity in visualizations, a pivotal feature noted in our study, to create more personalized, accessible and engaging experiences for readers.

Limitations

While this study has contributed valuable insights into the use of visualizations in CPM papers, there are limitations that warrant acknowledgment. The participant pool, primarily affiliated with academia and with substantial representation from North America, may not fully capture the diverse global perspectives and experiences of biomedical researchers more broadly. The insights gathered could be influenced by regional preferences and academic orientations, possibly limiting the generalizability of the findings. However, the preferred visualizations identified in this study may not encompass all potential visualizations or fully capture the nuanced preferences of different target users not covered in this study, such as clinicians. Acknowledging these limitations does not diminish the study's contributions but rather points to areas for refinement and exploration in future research.

However, it is essential to recognize that not all visualizations or methods should be followed uniformly across all CPM study reports. The selection and deployment of visualizations must be aligned with the specific research questions and the model at hand. Providing an exhaustive list of visualizations is not the intention here; rather, the goal is to demonstrate that effective visualization can be achieved by focusing on what is most relevant according to the main stakeholders in this field, biomedical researchers. This nuanced approach was based on three steps of Munzner's four-step process, providing a structured methodology that guides researchers in identifying and crafting visualizations that specifically suit their unique research context. From domain problem characterization to data/operation abstraction design and encoding/interaction technique design, Munzner's framework ensures that visualizations are developed with a clear understanding of the research goals and the target audience's needs. This alignment with Munzner's

principles may amplify the effectiveness of visualizations in CPM papers, enhancing their relevance, clarity, and impact.

6.6 CONCLUSIONS

In conclusion, this study provides a comprehensive finding of biomedical researchers' perspectives on the role of visualizations in ensuring that the results of CPM studies, especially preimpact analysis studies, are understandable, useful, and trustworthy. By examining the visualization preferences of authors and reviewers, the study not only underscores the crucial role of visualizations in presenting complex models but also emphasizes the importance of integrating both text and visuals for effective communication in CPM studies. The thorough mapping of participant preferences for visualizations across six stages of developing and validating CPM, including data exploration, predictor selection, modeling, result exploration, model performance, and model presentation, offers valuable insights into their practical application. Furthermore, cross-referencing between the visualization tasks and their corresponding themes across the three quality attributes provides a nuanced understanding of how the preferred visualizations can support CPM study results that are understandable, useful, and trustworthy. Therefore, this study demonstrated that involving biomedical researchers as authors, and reviewers might further ensure that CPM study results are understandable, useful, and trustworthy and advocates for a holistic approach to reporting CPM studies, where visualizations are not merely supplementary but integral to conveying complex information of CPM study results effectively.

Chapter 7. CONCLUSION AND CONTRIBUTION

This final Chapter summarizes the conclusions and contributions of my dissertation research, presented in Chapters 4 through 6, focusing on biomedical researchers. In Chapter 4, my study investigates the challenges biomedical researchers face in perceiving and producing CPM study results that meet the three quality attributes, understandable, useful, and trustworthy from the perspectives of authors and reviewers. I employed a mixed-method survey, using both quantitative and qualitative approaches in this Chapter. Chapter 5, based on interviews with a subset of survey respondents, outlines the specific needs required to address these challenges to ensure CPM study results meet the three quality attributes. Chapter 6, also based on those interviews, reports on biomedical researchers' preferences for visualizations in CPM studies, aiming to enhance the reporting quality of CPM study results across the three key quality attributes. Finally, I end this Chapter by describing contributions to the field of biomedical informatics, demonstrating the feasibility of involving a wide range of biomedical researchers to improve the quality of CPM study results, identifying gaps in the existing guideline, TRIPOD, and outlining potential avenues for future research.

7.1 SUMMARY OF FINDINGS

In Chapter 4, I comprehensively outline the challenges faced by biomedical researchers in ensuring understandable, useful, and trustworthy results in CPM studies. This comprehensive approach is facilitated by a mixed-method survey, which allows for a thorough exploration of the experiences of biomedical researchers as authors and reviewers of CPM study results reports. The respondents of this survey were predominantly over the age of 35, identified as men, and from academic institutions. Their experiences as authors and reviewers mostly involve CPM studies

focusing on preimpact analysis, which concerns the development and validation of models. Only one respondent reported experience as an author of impact analysis studies examining whether CPM impacts clinical practices using approaches such as RCTs.

This study uncovered biomedical researchers' challenges in perceiving and ensuring the qualities of CPM study results meet three quality attributes—understandable, useful, and trustworthy. 'Useful' emerges as the most challenging quality attribute, followed by 'trustworthy,' when respondents were asked about their perception of CPM study result qualities in research papers. As authors, biomedical researchers reported that achieving both 'useful' and 'trustworthy' in producing CPM study results has comparable challenges and is more challenging than 'understandable.' As reviewers, respondents consider providing feedback to ensure that CPM study results are 'understandable' to be as challenging as ensuring they are 'useful,' while regarded trustworthy study results seem to be more even challenging. These challenges are prevalent across different ranges of demographic characteristics and experiences, either as authors or reviewers, indicating that the challenges could be widespread among biomedical researchers.

The inclusion of qualitative analysis has deepened the understanding, revealing the reasons for the challenges researchers face in ensuring CPM results meet these three quality attributes as authors and reviewers. My study grouped these reasons into three groups of themes related to difficulties: 1) presenting understandable, useful, and trustworthy CPM study results, 2) conducting CPM studies, and 3) providing reviews to authors. The first group indicated that biomedical researchers acknowledge the difficulties in presenting CPM study results that meet the qualities across the three attributes. For 'understandable,' biomedical researchers acknowledge the challenge of effectively communicating complex CPM study results to a broader audience, particularly clinicians who are the target users of CPM. For 'useful,' biomedical researchers

emphasize the difficulties in designing CPM studies from the outset with clinical relevance in mind, aligning them with practical healthcare demands and involving collaboration among multiple stakeholders. For 'trustworthy,' biomedical researchers highlight challenges in ensuring trustworthy CPM study results, which require comprehensive reporting, including full disclosure of methodologies, data, and code, which can be challenging as it may reveal inconsistencies.

In Chapter 5, I describe the needs of biomedical researchers for results of CPM studies that are understandable, useful, and trustworthy. Through qualitative interviews and directed content analysis, a diverse array of needs were identified across these themes. Among the findings, several identified needs highlight their support for adherence to the reporting practices for preimpact analysis studies, such as the TRIPOD guidelines. The other findings also highlighted the need to adapt to advancements in the CPM field, such as integrating user-friendly tools for CPM applications, clarifying the contexts for CPM usage, and ensuring model equity across subgroups. Although many needs were unique to each quality attribute, overlaps were noted, especially with 'understandable' intersecting both 'useful' and 'trustworthy.' This suggests that ensuring CPM study results are understandable may concurrently address aspects of 'useful' and 'trustworthy.' For instance, when results are presented clearly and comprehensively, they are not only easier to comprehend (understandable) but also become more practical and actionable for clinical use (useful) and inspire greater confidence in their accuracy and integrity (trustworthy). However, comprehensive efforts are necessary to fully meet the needs across all three quality attributes. This involves implementing detailed writing for clear data presentation and interpretation, integrating user-friendly tools to enhance the practical application of CPMs, clarifying the specific contexts in which CPMs should be applied, and ensuring model equity across diverse population subgroups. By addressing these multifaceted needs, authors can present CPM study results that are not only

transparent and easy to understand but also highly applicable and reliable in various clinical settings thus improving the overall quality and impact of their research outcomes.

In Chapter 6, I report on the visualization preferences expressed by biomedical researchers to ensure that the presentations of CPM study results are understandable, useful, and trustworthy. This research offers an in-depth exploration of how biomedical researchers suggest visualizations can contribute to the effective communication of CPM study results, particularly for preimpact analysis studies. By identifying the preferences of biomedical researchers, the study highlights the pivotal role of visualizations in delineating complex data and the necessity of integrating text with visuals for reporting CPM study results. Extensive mapping of visualization preferences across the six stages of CPM development and validation—including data exploration, predictor selection, modeling, result exploration, model performance, and presentation (Lu, 2017; Steyerberg and Vergouwe, 2014)—provides crucial insights into their application.

Additionally, cross-referencing visualization tasks with the corresponding needs of biomedical researchers across the 'understandable,' 'useful,' and 'trustworthy' themes offers a nuanced perspective on how specific visualizations can meet these quality attributes. Each theme necessitates specific types of visualizations to ensure that CPM study results are understandable, useful, and trustworthy. For 'understandable' results, the focus is on clarity and accessibility, ensuring complex data and models are easily grasped by a wide audience. For 'useful' results, the emphasis is on practicality and user engagement, with interactive and user-friendly visualizations enhancing the application of CPMs in clinical workflows. Ensuring 'trustworthy' results demands accuracy and objectivity, supporting the credible presentation of data and findings. Thus, the study underscores the importance of engaging biomedical researchers to ensure that CPM study results are conveyed effectively, advocating for an integrated approach that uses both visualizations and

accompanying text to provide context in the dissemination of complex information within CPM study results.

In conclusion, this dissertation comprehensively examines the challenges, needs, and visualization preferences associated with ensuring understandable, useful, and trustworthy results in CPM studies, as perceived by biomedical researchers. From identifying the distinct challenges faced by biomedical researchers to outlining the specific needs for enhancing the quality of study results, and finally, to exploring the role of preferred visualizations in effectively communicating the study results, this research offers valuable insights into improving the reporting of CPM study results, specifically for preimpact analysis studies.

7.2 Contributions to Biomedical Informatics

Biomedical Informatics (BMI) is a field dedicated to promoting the effective use of biomedical data, information, and knowledge for scientific inquiry, problem-solving, and decision-making, with the aim of improving human health (Shortliffe and Cimino, 2014). This study makes an important contribution to BMI by identifying challenges in biomedical research experience with respect to three quality attributes of reporting CPM study results: understandable, useful, and trustworthy. These quality attributes represent the effective use of biomedical data in presenting results of CPM studies, aligning with the goals of BMI.

To the best of my knowledge, this study is the first to provide evidence that the biomedical research community recognizes these quality attributes as challenges in reporting CPM study results. While other quality attributes of CPM study reports, such as the risk of bias (as mentioned in PROBAS) and transparency (as noted in TRIPOD), are well acknowledged in systematic review studies (Collins et al., 2015), the quality attributes of understandable, useful, and trustworthy information are also important, as acknowledged by biomedical researchers in this study. This

acknowledgment validates the need for continued focus on enhancing the quality of these three quality attributes, efforts that will further contribute to the advancement of both BMI and the CPM field.

7.3 Contributions to demonstrating the feasibility of engaging Biomedical Researchers to improve the quality of CPM study results

This dissertation demonstrated a novel approach to engage biomedical researchers in improving the quality of reports in CPM studies. Previous initiatives to enhance CPM study reporting, such as TRIPOD, PROBAST, and SPIRIT AI, have relied on recruiting domain experts through professional networks and using Delphi methods to identify recommendations and achieve consensus on the recommendations for reporting in CPM studies (Moher et al., 2010). This traditional approach often excludes a broader range of biomedical researchers who may not be accessible via professional networks. Furthermore, this approach raised questions about whether these guidelines adequately represent the biomedical research community and address the challenges researchers face in achieving desired qualities in their studies. Delphi methods often focus directly on reaching consensus on the items to report in CPM studies.

My study recruited a broad range of biomedical researchers with experience as authors or reviewers of CPM studies through PubMed records (Chapter 3). First, my study focused on clarifying whether challenges in the three quality attributes of reporting CPM study results—understandable, useful, and trustworthy—exist among biomedical researchers, whether as authors, or reviewers, through a mixed-method survey (Chapter 4). Then, after confirming that these challenges exist, my study further identified needs and visualizations to ensure that CPM study results meet the three quality attributes through qualitative interviews (Chapters 5 and 6).

Therefore, my study demonstrated that recruiting a broad range of biomedical researchers to identify challenges and needs in presenting CPM study results can be an alternative approach to improving the reporting quality of CPM study results.

7.4 CONTRIBUTIONS TO IDENTIFYING GAPS IN TRIPOD

The needs and visualization preferences identified in Chapters 5 and 6, respectively, to ensure that CPM study results meet the three quality attributes—understandable, useful, and trustworthy—primarily pertain to the reporting of preimpact analysis in CPM studies. This focus aligns with the professional experience of study participants, who are primarily involved in authoring and reviewing such studies. Preimpact analysis studies are also the target of TRIPOD guidelines, which emphasize transparency as a key quality attribute. TRIPOD has issued its guidelines in two versions: the main guideline (Collins et al., 2015), and a second, more detailed version that includes explanations and elaborations (Moons et al., 2015). The latter provides a more comprehensive list of items to report, along with examples. These examples also feature visualizations for reporting results, such as figures for variable and participant selection, graphical scoring systems, nomograms, calibration plots, ROC curves, and net benefit plots.

However, by cross-referencing the findings in Chapter 5 and 6 with the TRIPOD guidelines, I was able to pinpoint specific areas where the gaps exist in the guideline. To identify these gaps, I cross-referenced both needs and visualization preferences with the TRIPOD guidelines and categorized them into three types: Type 1 includes needs and visualizations mentioned in TRIPOD; Type 2 comprises needs and visualizations not yet covered by TRIPOD; and Type 3 consists of needs and visualizations included in TRIPOD but requiring enhancements to the existing guidelines.

7.4.1 *Alignment and gap of biomedical researcher needs with TRIPOD*

I cross-referenced almost all needs with TRIPOD based on my discovery in Chapter 5, with one exception. This one exception lies in a need under the 'useful' theme and 'quality increase' sub-theme, where participants suggested RCTs to demonstrate that CPM can improve patient outcomes. Table 7.1 displays the alignment of each identified need with TRIPOD, categorized as Type 1 (covered by TRIPOD), Type 2 (not covered by TRIPOD), and Type 3 (covered by TRIPOD but requiring extension). The following paragraphs describe the needs within each theme.

Table 7.1 Mapping of the needs within each theme based on TRIPOD alignment

Theme	**Type**	**Sub-theme**	**# Needs**	**Total # Needs**
Understandable	Type 1	Knowledge	1-1-1	7
		Analysis	1-3-2, 1-3-5	
		Application	1-4-1, 1-4-2	
		Synthesis	1-5-1	
		Evaluation	1-6-1	
	Type 2	Comprehension	1-2-1	4
		Analysis	1-3-1, 1-3-3	
		Evaluation	1-6-2	
	Type 3	Analysis	1-3-4	2
		Evaluation	1-6-3	
Useful	Type 1	Easy to use	2-2-1	2
		Useful for tasks	2-5-1	
	Type 2	Critical support	2-1-1, 2-1-2	5
		More accomplishment	2-3-2	
		Quicker to complete	2-7-1	
		User-control	2-8-1	
	Type 3	More accomplishment	2-3-1	3
		Useful for tasks	2-5-2	
		Performance increase	2-6-1	
Trustworthy	Type 1	Objective	3-2-1, 3-2-2	6
		Validity	3-3-1, 3-3-2, 3-3-3, 3-3-4	
	Type 2	Accuracy	3-1-1	4
		Validity	3-3-5, 3-3-6	
		Stability	3-4-1	
	Type 3	Objective	3-2-3, 3-2-4	3

Theme	Type	Sub-theme	# Needs	Total # Needs
		Validity	3-3-7	

The needs within the 'understandable' theme exhibit a strong alignment with the TRIPOD guidelines, as evidenced by the predominance of Type 1 needs, which comprise approximately 54% (7/13) of the total needs under this theme. This high percentage suggests that TRIPOD effectively addresses many of the fundamental requirements for making CPM study results understandable. However, the remaining needs in this theme, categorized as Type 2 (30.8%, 4/13) and Type 3 (15.4%, 2/13), reveal areas where TRIPOD could improve. For instance, there is a clear call for more user-centric reporting to broaden the comprehensibility of study results, particularly under the 'comprehension' sub-theme (#Needs: 1-2-1). Participants also emphasized the necessity for a detailed breakdown of the data exploratory phase, focusing on initial data signals and specific data characteristics influential in model building and interpretation (#Needs: 1-3-1 and 1-3-3). Moreover, a multifaceted analysis of model performance using various metrics is suggested to enhance evaluation (#Needs: 1-6-2). These suggestions indicate that while TRIPOD provides a robust foundation, there is room for enhancement to fully address the complexities of making CPM study results understandable to a diverse audience.

The 'useful' theme presents a more varied alignment with TRIPOD, with Type 1 needs accounting for only 20% (2/10). This lower percentage compared to Type 2 (50%, 5/10) and Type 3 (30%, 3/10) underscores that TRIPOD covers some aspects of usefulness but leaves significant gaps. The emphasis in Type 2 needs is on guiding appropriate model use by clarifying functions and outlining limitations to avoid misuse (#Needs: 2-1-1), and cautioning against unintended consequences, including ethical concerns and healthcare burdens (#Needs: 2-1-2). These critical support elements are essential for practical application and are not sufficiently addressed by

TRIPOD. Furthermore, the 'more accomplishment' sub-theme suggests specific implementation strategies for practical model integration (#Needs: 2-3-2), highlighting the need for detailed guidelines. Enhancing content readability for quick comprehension is another priority under the 'quicker to complete' sub-theme, balancing concise text with visuals (#Need: 2-7-1). Additionally, the 'user-control' sub-theme underscores the importance of offering greater interactivity with results, such as interactive simulations (#Need: 2-8-1). Type 3 needs propose enhancements in TRIPOD, such as providing directions for future research on primary risk factors (#Needs: 2-3-1) and ensuring practical utility in clinical scenarios (#Needs: 2-5-2), which are crucial for making CPM results truly useful.

In the 'trustworthy' theme, Type 1 needs account for about 46% (6/13), highlighting a substantial alignment with TRIPOD. However, this still leaves a notable proportion of needs in Type 2 (30.8%, 4/13) and Type 3 (23.1%, 3/13), indicating areas where TRIPOD could be strengthened. Type 2 needs emphasize ensuring accuracy in data presentation to avoid discrepancies and misleading visual representations (#Needs: 3-1-1). This points to a critical gap in TRIPOD's current coverage, as accurate data presentation is essential for trustworthiness. Additionally, recognizing the limitations of established guidelines to maintain methodological soundness, using representative populations, and safeguarding against bias (#Needs: 3-3-5, 3-3-6) are emphasized under the 'validity' sub-theme. These needs highlight the importance of methodological rigor and representativeness in maintaining trustworthiness. Demonstrating model equity across different individuals and populations (#Needs: 3-4-1) under the 'stability' sub-theme is also crucial for ensuring that models are reliable and fair. Type 3 needs call for justifying research decisions, including analytical methods and variable selection, addressing anomalies, and substantiating model performance claims through multi-dimensional assessments and

benchmarking against existing models (#Needs: 3-2-3, 3-2-4). These enhancements underscore the need for comprehensive and trustworthy presentations of CPM study results.

This dissertation project identifies new needs that mostly indicate a requirement for more concerted efforts from authors to ensure that the quality of CPM studies is understandable, useful, and trustworthy, extending beyond TRIPOD's current recommendations. Thus, more effort is needed when authors aim to ensure their CPM study results meet these three quality attributes. This involves not only ensuring their report items adhere to existing guidelines in TRIPOD, which focuses on 'transparency' as its main quality attribute, but also verifying that their reported items fulfill the needs to make CPM study results understandable, useful, and trustworthy. By responding to these needs, authors can effectively tackle the complexities inherent in reporting CPM study results, thereby improving the quality of their research outcomes.

These findings also highlight that the distinct needs identified in this dissertation project, separate from TRIPOD and unique to each other as discussed in Chapter 5, support the notion that the three quality attributes—understandable, useful, and trustworthy—are critical and need to be considered when presenting study results in research papers. Additionally, these findings are rooted in historical perspectives that emphasize the distinctive features of these attributes, as introduced in Chapter 1. The 'understandable' attribute can be traced back to Bloom's Taxonomy of Educational Objectives, a significant work in the education field published in 1956, which was later expanded to assess students' comprehension of information (Bloom, 1956; Lord and Baviskar, 2007). The concept of 'useful' was first associated with perceived usefulness in management information systems in the early 1980s (Larcker and Lessig, 1980) and further developed in the Technology Acceptance Model (TAM) by Fred D. Davis in 1985 to include 'ease of use' and 'perceived usefulness' (Davis, 1985; Holden and Karsh, 2010; Venkatesh and Bala, 2008).

Regarding 'trustworthy,' Kelton's work in 2008 on Trust in Digital Information proposed a comprehensive model of trust in information, highlighting the importance of accuracy, objectivity, validity, and consistency in enhancing information dependability for decision-making (Kelton et al., 2008). These historical perspectives underscore the relevance and necessity of addressing these quality attributes—understandable, useful, and trustworthy—when presenting CPM study results.

7.4.2 *Alignment and gap in biomedical researchers' visualization preferences with TRIPOD*

To identify the alignment of visualization preferences with the visualizations presented in the second version of TRIPOD (Moons et al., 2015), I assessed each visualization task and its corresponding visual encoding identified in Chapter 6 and determined which sub-visualization tasks or visual encodings were covered in the TRIPOD guidelines and which were not. All the visualization tasks and visual encodings suggested by participants have their roots in the literature for reporting the development and validation of prediction models, as previously reported in Chapter 6.

Table 7.2 compiles the visual encodings according to visualization groups in reporting CPM study results (i.e., data exploration, feature selection, modeling, result exploration, model performance, and model presentation) and their corresponding alignment type with TRIPOD. Type 1 includes visual encoding already present in TRIPOD; Type 2 comprises visual encoding not yet included in TRIPOD. Type 3 covers those that are in TRIPOD but where participants highlighted additional visuals or features that could be added.

Table 7.2 Distributions of visualization tasks by their types

Type	**# Visual Encodings**
Type 1	Feature selections: 2-2, 2-3 Result exploration: 4-8 Model performance: 5-1, 5-2, 5-3

	Model presentation: 6-1 Multi-section: 7-1
Type 2	Data exploration: 1-1, 1-2, 1-3 Feature selection: 2-1 Modeling: 3-1 Result exploration: 4-1, 4-2, 4-3, 4-4, 4-6, 4-7 Model performance: 5-4, 5-5, 5-6, 5-7, 5-8, Model presentation: 6-2, 6-3
Type 3	Result exploration: 4-5 Model presentation: 6-4

As shown in Table 7.2., Type 2 visual encodings are predominant, indicating that participants suggested visualizations extending beyond those recommended in the existing TRIPOD guidelines. These encodings encompass a broad range of desired visual encodings, spreading across all six stages of developing and validating CPMs.

Type 1 visual encodings, which are in line with TRIPOD suggestions, are less numerous than Type 2. Notably, no visual encoding for data exploration and modeling is provided in TRIPOD.

For Type 3 visual encodings included in TRIPOD but where participants sought more detail, I observed a specific focus on result exploration and model presentation. In result exploration, a notable enhancement is the Restricted cubic spline, with a suggestion to show interactivity that illustrates how varying levels of continuous variables influence associated hazard or odds ratios. For the model presentation, participants expected to add interactivity, allowing for a more dynamic and engaging visualization experience.

The predominance of Type 2 visual encodings underscores a desire for a broader range of innovative visualizations, particularly in areas such as data exploration, result exploration, and model performance. Preferences expressed by participants suggest visualizations that extend beyond those currently recommended in the TRIPOD guidelines. This indicates the necessity for

continuous development and adaptation of visualization techniques to adeptly and quickly address the complexities of presenting CPM study results. These preferences suggest that while current TRIPOD guidelines provide a solid foundational framework, authors may benefit from incorporating a more diverse array of visualization strategies to capture better and convey intricate aspects of presenting CPM study results to readers.

Furthermore, my study, as reported specifically in Chapter 6, establishes groundwork for improving current reporting guidelines, specifically TRIPOD. By advocating for more detailed visualization tasks and encoding strategies from the perspective of biomedical researchers, my findings can inform the communication of results in CPM studies to be more understandable, useful, and trustworthy. My findings also highlight the potential benefits of integrating Human-Centered Design principles (FDIs, 2009) and Munzner's framework (Munzner, 2009) to refine the use of visualizations in CPM study results.

The findings from my dissertation highlight the specific needs and visualization preferences identified by biomedical researchers to enhance the quality of CPM study results, suggesting potential updates to the existing TRIPOD guidelines. These needs, although not directly warranting an immediate update to the guidelines, underscore the importance of incorporating firsthand insights from active researchers in the field. By aligning the TRIPOD guidelines more closely with the real-world challenges faced by those who author and review CPM studies, this approach ensures that the guidelines remain relevant and practical. This alignment not only supports the current standards but also enriches them, making the guidelines more comprehensive and responsive to the evolving demands of CPM study reporting.

7.5 CONTRIBUTIONS TO MOVING PREIMPACT ANALYSIS STUDIES FORWARD

My study's findings have identified needs and visualization preferences from the perspective of biomedical researchers to ensure that the results of CPM studies are understandable, useful, and trustworthy. These needs and visualization preferences extend beyond those suggested by existing guidelines for preimpact analysis CPM studies, such as TRIPOD (Collins et al., 2015). Thus, these findings suggest that the three quality attributes may complement the existing quality attributes of transparency and completeness, as suggested by TRIPOD, in the effort to improve the quality of CPM study reporting. Frameworks such as Bloom's Taxonomy for understandability, the Technology Acceptance Model (TAM) for usefulness, and Trust in Digital for trustworthiness suggest that possessing these qualities can encourage target users to take follow-up actions. Therefore, addressing these three quality attributes in CPM study reports may encourage biomedical researchers to engage with preimpact analysis study results and undertake follow-up studies. However, my study also noted challenges with the quality of the preimpact analysis studies across the three quality attributes.

In Chapter 4, my study found that biomedical researchers who responded to and participated in the survey predominantly focused on the initial stages of CPMs, particularly preimpact analysis studies. This focus contrasts with the recruitment efforts described in Chapter 3, which aimed to involve all biomedical researchers with experience as authors and reviewers of CPM studies in research papers, regardless of the type of CPM study. My participants, predominantly biomedical researchers, have more experience in pre-impact analysis studies than in impact analysis, somewhat supporting Keogh et al., (Keogh et al., 2014), which suggests that impact analysis is more prevalent. However, this should be viewed with caution due to my study's reliance on convenience sampling. Moreover, previous studies also show a tendency to develop

redundant CPMs for similar clinical issues without progressing to more critical stages like external validation and impact analysis, exacerbating this issue (Binuya et al., 2022; Damen et al., 2016; Phung et al., 2019).

As the current state of practices in CPM studies highlights that most such studies in the field are preimpact analysis studies, there should be a shift toward follow-up studies that require more involvement of clinicians to move CPMs closer to use in clinical practices. These follow-up studies could include impact analysis studies such as randomized controlled trials (RCTs) or human-centered design studies aimed at developing clinical decision support tools using CPMs that engage clinicians as the target users. Therefore, by incorporating the additional three quality attributes—understandable, useful, and trustworthy—in reporting preimpact analyses study results, I aim to contribute to promoting the uptake of these CPM studies. For biomedical researchers serving as authors and reviewers, using the identified needs and visualization preferences can aid in reporting CPM studies that meet the three quality attributes. For biomedical researchers, these needs and visualization preferences can guide them in assessing which CPM study results meet these three quality attributes and then select them for follow-up studies.

7.6 FUTURE STUDIES

Despite the comprehensive findings from my dissertation, further studies are needed to incorporate the needs of broader target users and improve the quality of CPM studies. This study focused only on one group of target users, biomedical researchers; therefore, similar studies involving other target groups, such as clinicians, are warranted. These investigations may engage clinicians as key target users to identify their challenges, needs, and visualization preferences for improving the quality of preimpact analysis studies in research papers across the three quality attributes: understandable, useful, and trustworthy. This approach is informed by the recognition,

both within this dissertation and supported by other literature, that clinicians are not only the target users of preimpact analysis studies but may also be involved in developing CPMs (Strandberg et al., 2024). Thus, involving clinicians in such studies would ensure that the reporting of CPM study results aligns with their professional requirements.

Subsequent studies could leverage the identified needs for understandable, useful, and trustworthy results of CPM studies from both biomedical researchers and clinicians to evaluate the quality of existing CPM research papers. This assessment could be carried out through systematic or scoping review studies, critically analyzing how the current reporting of CPM study results aligns with these identified needs, thereby providing a comprehensive overview of the state of CPM reporting practices.

The other future study could investigate incorporating the needs of the target users into the reporting of CPM study results, leading to measurable outcomes among readers. For example, future research could compare CPM study results that incorporate identified needs against standard practices. Subsequently, it could measure whether readers perceive improvements in the three quality attributes—understandable, useful, and trustworthy—using standardized metrics between the two approaches.

BIBLIOGRAPHY

Ahuja, M., Jr, R.A., 2021. Barriers to Dissemination of Local Health Data Faced by US State Agencies: Survey Study of Behavioral Risk Factor Surveillance System Coordinators. J. Med. Internet Res. 23, e16750. https://doi.org/10.2196/16750

Alharbey, R., Dessouky, M., Sedik, A., Siam, A., Elaskily, M., 2022. Fatigue State Detection for Tired Persons in Presence of Driving Periods. IEEE Access 10, 1–1. https://doi.org/10.1109/ACCESS.2022.3185251

An, C., Oh, H., Chang, J., Oh, S.-J., Lee, J., Han, C., Kim, S., 2021. Development and validation of a prognostic model for early triage of patients diagnosed with COVID-19. Sci. Rep. 11. https://doi.org/10.1038/s41598-021-01452-7

Ancker, J.S., Benda, N.C., Reddy, M., Unertl, K.M., Veinot, T., 2021. Guidance for publishing qualitative research in informatics. J. Am. Med. Inform. Assoc. JAMIA 28, 2743–2748. https://doi.org/10.1093/jamia/ocab195

Anderson, L.W., Krathwohl, D.R., 2001. A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives. Longman,.

Antes, A.L., Kuykendall, A., DuBois, J.M., 2019. The lab management practices of "Research Exemplars" that foster research rigor and regulatory compliance: A qualitative study of successful principal investigators. PLOS ONE 14, e0214595. https://doi.org/10.1371/journal.pone.0214595

Arslan, J., Benke, K., 2023. Statistical Analysis of Ceiling and Floor Effects in Medical Trials. Appl. Biosci. 2, 668–681. https://doi.org/10.3390/applbiosci2040042

Authorship in MEDLINE [WWW Document], 2023. URL https://www.nlm.nih.gov/bsd/policy/authorship.html (accessed 4.2.24).

Balogh, E.P., Miller, B.T., Ball, J.R., Care, C. on D.E. in H., Services, B. on H.C., Medicine, I. of, The National Academies of Sciences, E., 2015. Overview of Diagnostic Error in Health Care, Improving Diagnosis in Health Care. National Academies Press (US).

Barda, A.J., Horvat, C.M., Hochheiser, H., 2020. A qualitative research framework for the design of user-centered displays of explanations for machine learning model predictions in healthcare. BMC Med. Inform. Decis. Mak. 20, 257. https://doi.org/10.1186/s12911-020-01276-x

Barnett, M.L., Boddupalli, D., Nundy, S., Bates, D.W., 2019. Comparative Accuracy of Diagnosis by Collective Intelligence of Multiple Physicians vs Individual Physicians. JAMA Netw. Open 2, e190096–e190096. https://doi.org/10.1001/jamanetworkopen.2019.0096

Bellón, J.Á., de Dios Luna, J., King, M., Nazareth, I., Motrico, E., GildeGómez-Barragán, M.J., Torres-González, F., Montón-Franco, C., Sánchez-Celaya, M., Díaz-Barreiros, M.Á., Vicens, C., Moreno-Peral, P., 2017. Predicting the onset of hazardous alcohol drinking in primary care: development and validation of a simple risk algorithm. Br. J. Gen. Pract. J. R. Coll. Gen. Pract. 67, e280–e292. https://doi.org/10.3399/bjgp17X690245

Berner, E.S., Graber, M.L., 2008. Overconfidence as a cause of diagnostic error in medicine. Am. J. Med. 121, S2-23. https://doi.org/10.1016/j.amjmed.2008.01.001

Binuya, M.A.E., Engelhardt, E.G., Schats, W., Schmidt, M.K., Steyerberg, E.W., 2022. Methodological guidance for the evaluation and updating of clinical prediction models: a systematic review. BMC Med. Res. Methodol. 22, 316. https://doi.org/10.1186/s12874-022-01801-8

Bloom, B.S., 1956. Taxonomy of educational objectives: The classification of educational goals.
Bonnett, L.J., Snell, K.I.E., Collins, G.S., Riley, R.D., 2019. Guide to presenting clinical prediction models for use in clinical settings. BMJ 365, l737. https://doi.org/10.1136/bmj.l737
Bösner, S., Haasenritter, J., Becker, A., Karatolios, K., Vaucher, P., Gencer, B., Herzig, L., Heinzel-Gutenbrunner, M., Schaefer, J.R., Hani, M.A., 2010. Ruling out coronary artery disease in primary care: development and validation of a simple prediction rule. Cmaj 182, 1295–1300.
Braun, V., Clarke, V., 2006. Using thematic analysis in psychology. Qual. Res. Psychol. 3, 77–101. https://doi.org/10.1191/1478088706qp063oa
Briggs, A.H., Weinstein, M.C., Fenwick, E.A.L., Karnon, J., Sculpher, M.J., Paltiel, A.D., 2012. Model Parameter Estimation and Uncertainty: A Report of the ISPOR-SMDM Modeling Good Research Practices Task Force-6. Value Health 15, 835–842. https://doi.org/10.1016/j.jval.2012.04.014
Brown, S., 2010. Likert scale examples for surveys.
Bruce, P., Bruce, A., 2017. Practical Statistics for Data Scientists: 50 Essential Concepts, 1st edition. ed. O'Reilly Media, Sebastopol, CA.
Burns, A., Xiong, C., Franconeri, S., Cairo, A., Mahyar, N., 2020. How to evaluate data visualizations across different levels of understanding.
Campbell, P., Bishop, A., Dunn, K.M., Main, C.J., Thomas, E., Foster, N.E., 2013. Conceptual overlap of psychological constructs in low back pain. Pain 154, 1783–1791. https://doi.org/10.1016/j.pain.2013.05.035
Centor, R.M.., Witherspoon, J.M., Dalton, H.P., Brody, C.E., Link, K., 1981. The Diagnosis of Strep Throat in Adults in the Emergency Room. Med. Decis. Making 1, 239–246. https://doi.org/10.1177/0272989X8100100304
Chatzimparmpas, A., Martins, R.M., Jusufi, I., Kucher, K., Rossi, F., Kerren, A., 2020. The State of the Art in Enhancing Trust in Machine Learning Models with the Use of Visualizations. Comput. Graph. Forum 39, 713–756. https://doi.org/10.1111/cgf.14034
Chiang, S., Moss, R., Black, A.P., Jackson, M., Moss, C., Bidwell, J., Meisel, C., Loddenkemper, T., 2021. Evaluation and recommendations for effective data visualization for seizure forecasting algorithms. JAMIA Open 4, ooab009. https://doi.org/10.1093/jamiaopen/ooab009
Collins, G.S., Altman, D.G., 2010. An independent and external validation of QRISK2 cardiovascular disease risk score: a prospective open cohort study. BMJ 340. https://doi.org/10.1136/bmj.c2442
Collins, G.S., Reitsma, J.B., Altman, D.G., Moons, K.G., 2015. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD Statement. BMC Med. 13, 1. https://doi.org/10.1186/s12916-014-0241-z
Corazzini, J.G., 1977. Trust as a complex multi-dimensional construct. Psychol. Rep. 40, 75–80. https://doi.org/10.2466/pr0.1977.40.1.75
Cowley, L.E., Farewell, D.M., Maguire, S., Kemp, A.M., 2019. Methodological standards for the development and evaluation of clinical prediction rules: a review of the literature. Diagn. Progn. Res. 3, 16. https://doi.org/10.1186/s41512-019-0060-y
cran, r, n.d. The Comprehensive R Archive Network [WWW Document]. URL https://cran.r-project.org/ (accessed 1.13.24).

Create a mail merge with Gmail & Google Sheets | Apps Script [WWW Document], n.d. . Google Dev. URL https://developers.google.com/apps-script/samples/automations/mail-merge (accessed 3.1.23).

Cruz Rivera, S., Liu, X., Chan, A.-W., Denniston, A.K., Calvert, M.J., 2020. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. Nat. Med. 26, 1351–1363. https://doi.org/10.1038/s41591-020-1037-7

Cui, W., 2019. Visual Analytics: A Comprehensive Overview. IEEE Access 1–1. https://doi.org/10.1109/ACCESS.2019.2923736

Damen, J.A.A.G., Hooft, L., Schuit, E., Debray, T.P.A., Collins, G.S., Tzoulaki, I., Lassale, C.M., Siontis, G.C.M., Chiocchia, V., Roberts, C., Schlüssel, M.M., Gerry, S., Black, J.A., Heus, P., Schouw, Y.T. van der, Peelen, L.M., Moons, K.G.M., 2016. Prediction models for cardiovascular disease risk in the general population: systematic review. BMJ 353, i2416. https://doi.org/10.1136/bmj.i2416

Davies, K., 2007. The information-seeking behaviour of doctors: a review of the evidence. Health Inf. Libr. J. 24, 78–94. https://doi.org/10.1111/j.1471-1842.2007.00713.x

Davis, F.D., 1989. Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Q. 319–340.

Davis, F.D., 1985. A technology acceptance model for empirically testing new end-user information systems: Theory and results (PhD Thesis). Massachusetts Institute of Technology.

Devanarayan, V., Ye, Y., Charil, A., Andreozzi, E., Sachdev, P., Llano, D.A., Tian, L., Zhu, L., Hampel, H., Kramer, L., Dhadda, S., Irizarry, M., Initiative (ADNI), for the A.D.N., 2023. Predicting clinical progression trajectories of early Alzheimer's disease patients. Alzheimers Dement. n/a. https://doi.org/10.1002/alz.13565

Dhiman, P., Ma, J., Navarro, C.A., Speich, B., Bullock, G., Damen, J.A., Kirtley, S., Hooft, L., Riley, R.D., Van Calster, B., Moons, K.G.M., Collins, G.S., 2021. Reporting of prognostic clinical prediction models based on machine learning methods in oncology needs to be improved. J. Clin. Epidemiol. 138, 60–72. https://doi.org/10.1016/j.jclinepi.2021.06.024

Donaldson, D.R., 2016. The Digitized Archival Document Trustworthiness Scale. https://doi.org/10.2218/IJDC.V11I1.387

Ebell, M.H., Afonso, A.M., Gonzales, R., Stein, J., Genton, B., Senn, N., 2012. Development and validation of a clinical decision rule for the diagnosis of influenza. J. Am. Board Fam. Med. JABFM 25, 55–62. https://doi.org/10.3122/jabfm.2012.01.110161

Ebell, M.H., Rahmatullah, I., Cai, X., Bentivegna, M., Hulme, C., Thompson, M., Lutz, B., 2021. A Systematic Review of Clinical Prediction Rules for the Diagnosis of Influenza. J. Am. Board Fam. Med. 34, 1123–1140. https://doi.org/10.3122/jabfm.2021.06.210110

Eberhard, K., 2023. The effects of visualization on judgment and decision-making: a systematic literature review. Manag. Rev. Q. 73, 167–214. https://doi.org/10.1007/s11301-021-00235-8

Ely, J.W., Graber, M.L., 2016. Preventing Diagnostic Errors in Primary Care. Am. Fam. Physician 94, 426–432.

ESHRE Capri Workshop Group, 2018. Protect us from poor-quality medical research. Hum. Reprod. 33, 770–776. https://doi.org/10.1093/humrep/dey056

Faieta, J., Bourassa, J., Best, K., 2024. Refinement of Health App Review Tool (HART) through stakeholder interviews: HART 2.0. Assist. Technol. 36, 75–81. https://doi.org/10.1080/10400435.2023.2213742

FDIs, I., 2009. 9241-210: 2009. Ergonomics of human system interaction-Part 210: Human-centered design for interactive systems (formerly known as 13407). Int. Organ. Stand. ISO Switz.

Fosnacht, K., Sarraf, S., Howe, E., Peck, L.K., 2017. How Important are High Response Rates for College Surveys? Rev. High. Educ. 40, 245–265. https://doi.org/10.1353/rhe.2017.0003

Gage, B.F., Waterman, A.D., Shannon, W., Boechler, M., Rich, M.W., Radford, M.J., 2001. Validation of clinical classification schemes for predicting stroke: results from the National Registry of Atrial Fibrillation. JAMA 285, 2864–2870. https://doi.org/10.1001/jama.285.22.2864

Gail, M.H., Brinton, L.A., Byar, D.P., Corle, D.K., Green, S.B., Schairer, C., Mulvihill, J.J., 1989. Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. J. Natl. Cancer Inst. 81, 1879–1886. https://doi.org/10.1093/jnci/81.24.1879

Gambetta, D., 2000. Can We Trust Trust? Diego Gambetta.

Ge, M., 2009. Information quality assessment and effects on inventory decision-making (PhD Thesis). Dublin City University.

George, E., 2020. Trust in peer review, from the perspective of peer reviewers (Part 1) [WWW Document]. Ed. Insights. URL https://www.editage.com/insights/trust-in-peer-review-from-the-perspective-of-peer-reviewers-part-1 (accessed 1.21.24).

Gotz, D., Borland, D., 2016. Data-Driven Healthcare: Challenges and Opportunities for Interactive Visualization. IEEE Comput. Graph. Appl. 36, 90–96. https://doi.org/10.1109/MCG.2016.59

Graber, M.L., 2013. The incidence of diagnostic error in medicine. BMJ Qual. Saf. 22 Suppl 2, ii21–ii27. https://doi.org/10.1136/bmjqs-2012-001615

Haasenritter, J., Donner-Banzhoff, N., Bösner, S., 2015. Chest pain for coronary heart disease in general practice: clinical judgement and a clinical decision rule. Br. J. Gen. Pract. J. R. Coll. Gen. Pract. 65, e748-753. https://doi.org/10.3399/bjgp15X687385

Hamilton, W., 2009. The CAPER studies: five case-control studies aimed at identifying and quantifying the risk of cancer in symptomatic primary care patients. Br. J. Cancer 101 Suppl 2, S80-86. https://doi.org/10.1038/sj.bjc.6605396

Han, P.K.J., Klein, W.M.P., Lehman, T.C., Massett, H., Lee, S.C., Freedman, A.N., 2009. Laypersons' responses to the communication of uncertainty regarding cancer risk estimates. Med. Decis. Mak. Int. J. Soc. Med. Decis. Mak. 29, 391–403. https://doi.org/10.1177/0272989X08327396

Haskins, R., Osmotherly, P.G., Tuyl, F., Rivett, D.A., 2013. Uncertainty in Clinical Prediction Rules: The Value of Credible Intervals. J. Orthop. Sports Phys. Ther. 44, 85–91. https://doi.org/10.2519/jospt.2014.4877

Haybittle, J.L., Blamey, R.W., Elston, C.W., Johnson, J., Doyle, P.J., Campbell, F.C., Nicholson, R.I., Griffiths, K., 1982. A prognostic index in primary breast cancer. Br. J. Cancer 45, 361–366.

Heckerling, P.S., Tape, T.G., Wigton, R.S., Hissong, K.K., Leikin, J.B., Ornato, J.P., Cameron, J.L., Racht, E.M., 1990. Clinical prediction rule for pulmonary infiltrates. Ann. Intern. Med. 113, 664–670. https://doi.org/10.7326/0003-4819-113-9-664

Hemming, K., Taljaard, M., 2021. Knowledge translation of prediction rules: methods to help health professionals understand their trade-offs. Diagn. Progn. Res. 5, 21. https://doi.org/10.1186/s41512-021-00109-3

Heus, P., Damen, J.A.A.G., Pajouheshnia, R., Scholten, R.J.P.M., Reitsma, J.B., Collins, G.S., Altman, D.G., Moons, K.G.M., Hooft, L., 2018. Poor reporting of multivariable prediction model studies: towards a targeted implementation strategy of the TRIPOD statement. BMC Med. 16, 120. https://doi.org/10.1186/s12916-018-1099-2

Hippisley-Cox, J., Coupland, C., 2012. Identifying patients with suspected colorectal cancer in primary care: derivation and validation of an algorithm. Br. J. Gen. Pract. 62, e29–e37.

Holden, R.J., Karsh, B.-T., 2010. The Technology Acceptance Model: Its past and its future in health care. J. Biomed. Inform. 43, 159–172. https://doi.org/10.1016/j.jbi.2009.07.002

Horowitz, Noya, Moshkowitz, M., Halpern, Z., Leshno, M., 2007. Applying data mining techniques in the development of a diagnostics questionnaire for GERD. Dig. Dis. Sci. 52, 1871–1878.

Horowitz, N., Moshkowitz, M., Leshno, M., Ribak, J., Birkenfeld, S., Kenet, G., Halpern, Z., 2007. Clinical trial: evaluation of a clinical decision-support model for upper abdominal complaints in primary-care practice. Aliment. Pharmacol. Ther. 26, 1277–1283. https://doi.org/10.1111/j.1365-2036.2007.03497.x

Hsieh, H.-F., Shannon, S.E., 2005. Three approaches to qualitative content analysis. Qual. Health Res. 15, 1277–1288. https://doi.org/10.1177/1049732305276687

Hussain, M.I., Figueiredo, M.C., Tran, B.D., Su, Z., Molldrem, S., Eikey, E.V., Chen, Y., 2020. A scoping review of qualitative research in JAMIA: past contributions and opportunities for future work. J. Am. Med. Inform. Assoc. JAMIA 28, 402–413. https://doi.org/10.1093/jamia/ocaa179

Hyun, S., Johnson, S.B., Stetson, P.D., Bakken, S., 2009. Development and evaluation of nursing user interface screens using multiple methods. J. Biomed. Inform. 42, 1004–1012. https://doi.org/10.1016/j.jbi.2009.05.005

IEC, 2011. ISO-IEC 25010: 2011 Systems and software engineering - Systems and software Quality Requirements and Evaluation (SQuaRE) - System and software quality models. ISO, Geneva.

ITHS, n.d. Primary Care Research Network. ITHS. URL https://www.iths.org/investigators/find-collaborators/primary-care-research-network/ (accessed 1.14.24).

Jadad, A.R., Gagliardi, A., 1998. Rating health information on the Internet: navigating to knowledge or to Babel? JAMA 279, 611–614. https://doi.org/10.1001/jama.279.8.611

Jensen, J.D., King, A.J., Carcioppolo, N., Davis, L., 2012. Why are Tailored Messages More Effective? A Multiple Mediation Analysis of a Breast Cancer Screening Intervention. J. Commun. 62, 851–868. https://doi.org/10.1111/j.1460-2466.2012.01668.x

Kappen, T.H., van Loon, K., Kappen, M.A.M., van Wolfswinkel, L., Vergouwe, Y., van Klei, W.A., Moons, K.G.M., Kalkman, C.J., 2016. Barriers and facilitators perceived by physicians when using prediction models in practice. J. Clin. Epidemiol. 70, 136–145. https://doi.org/10.1016/j.jclinepi.2015.09.008

Kareemi, H., Vaillancourt, C., Rosenberg, H., Fournier, K., Yadav, K., 2021. Machine Learning Versus Usual Care for Diagnostic and Prognostic Prediction in the Emergency

Department: A Systematic Review. Acad. Emerg. Med. 28, 184–196. https://doi.org/10.1111/acem.14190
Karhade, Aditya.V., Schwab, J.H., 2020. CORR Synthesis: When Should We Be Skeptical of Clinical Prediction Models? Clin. Orthop. 478, 2722–2728. https://doi.org/10.1097/CORR.0000000000001367
Karlin-Zysman, C., Zeitoun, N., Belletti, L., McCullagh, L., McGinn, T., 2012. Struggling to bring clinical prediction rules to the point of care: missed opportunities to impact patient care. J. Comp. Eff. Res. 1, 421–429. https://doi.org/10.2217/cer.12.51
Keim, D., Andrienko, G., Fekete, J.-D., Görg, C., Kohlhammer, J., Melançon, G., 2008. Visual analytics: Definition, process, and challenges, in: Information Visualization. Springer, pp. 154–175.
Kelton, K., Fleischmann, K.R., Wallace, W.A., 2008. Trust in digital information. J. Am. Soc. Inf. Sci. Technol. 59, 363–374. https://doi.org/10.1002/asi.20722
Kennedy, G., Gallego, B., 2019. Clinical prediction rules: A systematic review of healthcare provider opinions and preferences. Int. J. Med. Inf. 123, 1–10. https://doi.org/10.1016/j.ijmedinf.2018.12.003
Keogh, C., Wallace, E., O'Brien, K.K., Galvin, R., Smith, S.M., Lewis, C., Cummins, A., Cousins, G., Dimitrov, B.D., Fahey, T., 2014. Developing an international register of clinical prediction rules for use in primary care: a descriptive analysis. Ann. Fam. Med. 12, 359–366. https://doi.org/10.1370/afm.1640
Kerracher, N., Kennedy, J., 2017. Constructing and Evaluating Visualisation Task Classifications: Process and Considerations. Comput. Graph. Forum 36, 47–59. https://doi.org/10.1111/cgf.13167
Keylen, P. van der, Tomandl, J., Wollmann, K., Möhler, R., Sofroniou, M., Maun, A., Voigt-Radloff, S., Frank, L., 2020. The Online Health Information Needs of Family Physicians: Systematic Review of Qualitative and Quantitative Studies. J. Med. Internet Res. 22, e18816. https://doi.org/10.2196/18816
Khalifa, M., 2019. Using PubMed to Generate Email Lists of Participants for Healthcare Survey Research: A Simple and Practical Approach. Stud. Health Technol. Inform. 262, 348–351. https://doi.org/10.3233/SHTI190090
Khullar, D., Jena, A.B., 2016. Reducing prognostic errors: a new imperative in quality healthcare. BMJ 352. https://doi.org/10.1136/bmj.i1417
King, M., Marston, L., Švab, I., Maaroos, H.-I., Geerlings, M.I., Xavier, M., Benjamin, V., Torres-Gonzalez, F., Bellon-Saameno, J.A., Rotar, D., 2011. Development and validation of a risk model for prediction of hazardous alcohol consumption in general practice attendees: the predictAL study. PLoS One 6, e22175.
Klement, W., El Emam, K., 2023. Consolidated Reporting Guidelines for Prognostic and Diagnostic Machine Learning Modeling Studies: Development and Validation. J. Med. Internet Res. 25, e48763. https://doi.org/10.2196/48763
Knowles, S.E., Ercia, A., Caskey, F., Rees, M., Farrington, K., Van der Veer, S.N., 2021. Participatory co-design and normalisation process theory with staff and patients to implement digital ways of working into routine care: the example of electronic patient-reported outcomes in UK renal services. BMC Health Serv. Res. 21, 706. https://doi.org/10.1186/s12913-021-06702-y
Kostopoulou, O., Rosen, A., Round, T., Wright, E., Douiri, A., Delaney, B., 2015. Early diagnostic suggestions improve accuracy of GPs: a randomised controlled trial using

computer-simulated patients. Br. J. Gen. Pract. J. R. Coll. Gen. Pract. 65, e49-54. https://doi.org/10.3399/bjgp15X683161

Lallemand, C., Koenig, V., 2020. Measuring the Contextual Dimension of User Experience: Development of the User Experience Context Scale (UXCS), in: Proceedings of the 11th Nordic Conference on Human-Computer Interaction: Shaping Experiences, Shaping Society. Presented at the NordiCHI '20: Shaping Experiences, Shaping Society, ACM, Tallinn Estonia, pp. 1–13. https://doi.org/10.1145/3419249.3420156

Lampe, K., Doupi, P., Hoven, J.M. van den, 2003. Internet Health Resources: from Quality to Trust. Methods Inf. Med. 42, 134–142. https://doi.org/10.1055/s-0038-1634324

Laposata, M., 2018. The Definition and Scope of Diagnostic Error in the US and How Diagnostic Error is Enabled. J. Appl. Lab. Med. 3, 128–134. https://doi.org/10.1373/jalm.2017.025882

Larcker, D., Lessig, V., 2007. Perceived Usefulness of Information: A Psychometric Examination. Decis. Sci. 11, 121–134. https://doi.org/10.1111/j.1540-5915.1980.tb01130.x

Larcker, D.F., Lessig, V.P., 1980. Perceived Usefulness of Information: A Psychometric Examination*. Decis. Sci. 11, 121–134. https://doi.org/10.1111/j.1540-5915.1980.tb01130.x

Lau, F., Price, M., 2017. Chapter 3 Clinical Adoption Framework, in: Handbook of eHealth Evaluation: An Evidence-Based Approach [Internet]. University of Victoria.

Lenney, E., 1977. Women's self-confidence in achievement settings. Psychol. Bull. 84, 1–13. https://doi.org/10.1037/0033-2909.84.1.1

Lester, M.A., 2022. Using the Community of Practice Framework to Examine Informal Science Educators' Epistemological and Pedagogical Beliefs in Informal Science.

Lim, W.S., van der Eerden, M.M., Laing, R., Boersma, W.G., Karalus, N., Town, G.I., Lewis, S.A., Macfarlane, J.T., 2003. Defining community acquired pneumonia severity on presentation to hospital: an international derivation and validation study. Thorax 58, 377–382. https://doi.org/10.1136/thorax.58.5.377

Liu, J., Dan, W., Liu, X., Zhong, X., Chen, C., He, Q., Wang, J., 2023. Development and validation of predictive model based on deep learning method for classification of dyslipidemia in Chinese medicine. Health Inf. Sci. Syst. 11, 21. https://doi.org/10.1007/s13755-023-00215-0

Lor, M., Koleck, T.A., Bakken, S., 2019. Information visualizations of symptom information for patients and providers: a systematic review. J. Am. Med. Inform. Assoc. 26, 162–171. https://doi.org/10.1093/jamia/ocy152

Lord, T., Baviskar, S., 2007. Moving Students From Information Recitation to Information Understanding: Exploiting Bloom's Taxonomy in Creating Science Questions. J. Coll. Sci. Teach. 36, 40–44.

Lu, Y., 2017. Methodologies in Predictive Visual Analytics (PhD Thesis). Arizona State University.

Lüdecke, D., 2023. Plotting Interaction Effects of Regression Models [WWW Document]. URL https://cran.r-project.org/web/packages/sjPlot/vignettes/plot_interactions.html (accessed 8.6.23).

Luo, W., Phung, D., Tran, T., Gupta, S., Rana, S., Karmakar, C., Shilton, A., Yearwood, J., Dimitrova, N., Ho, T.B., Venkatesh, S., Berk, M., 2016. Guidelines for Developing and

Reporting Machine Learning Predictive Models in Biomedical Research: A Multidisciplinary View. J. Med. Internet Res. 18, e323. https://doi.org/10.2196/jmir.5870
Ma, K.-L., Liao, I., Frazier, J., Hauser, H., Kostis, H.-N., 2012. Scientific Storytelling Using Visualization. IEEE Comput. Graph. Appl. 32, 12–19. https://doi.org/10.1109/MCG.2012.24
Maguire, M., 2001. Methods to support human-centred design. Int. J. Hum.-Comput. Stud. 55, 587–634. https://doi.org/10.1006/ijhc.2001.0503
Mahyar, N., Kim, S.-H., Kwon, B.C., 2015. Towards a taxonomy for evaluating user engagement in information visualization, in: Workshop on Personal Visualization: Exploring Everyday Life.
Maiga, A., Farjah, F., Blume, J., Deppen, S., Welty, V.F., D'Agostino, R.S., Colditz, G.A., Kozower, B.D., Grogan, E.L., 2019. Risk Prediction in Clinical Practice—A Practical Guide for Cardiothoracic Surgeons. Ann. Thorac. Surg. 108, 1573–1582. https://doi.org/10.1016/j.athoracsur.2019.04.126
Mangiafico, S., 2016. R Handbook: Friedman Test [WWW Document]. URL https://rcompanion.org/handbook/F_10.html (accessed 11.6.23).
Marshall, E., Marquier, B., n.d. Friedman in R.
McGinn, T.G., Deluca, J., Ahlawat, S.K., Mobo, B.H., Wisnivesky, J.P., 2003. Validation and modification of streptococcal pharyngitis clinical prediction rules. Mayo Clin. Proc. 78, 289–293. https://doi.org/10.4065/78.3.289
McGuinness, B., Leggatt, A., 2006. Information Trust and Distrust in a Sensemaking Task.
Medic, G., Kosaner Kließ, M., Atallah, L., Weichert, J., Panda, S., Postma, M., EL-Kerdi, A., 2019. Evidence-based Clinical Decision Support Systems for the prediction and detection of three disease states in critical care: A systematic literature review. F1000Research 8, 1728. https://doi.org/10.12688/f1000research.20498.2
MEDLINE/PubMed Data Element (Field) Descriptions [WWW Document], n.d. URL https://www.nlm.nih.gov/bsd/mms/medlineelements.html (accessed 3.1.23).
Meeßen, S.M., Thielsch, M.T., Hertel, G., 2020. Trust in Management Information Systems (MIS). Z. Für Arb.- Organ. AO 64, 6–16. https://doi.org/10.1026/0932-4089/a000306
Meyer, V.M., Benjamens, S., Moumni, M.E., Lange, J.F.M., Pol, R.A., 2022. Global Overview of Response Rates in Patient and Health Care Professional Surveys in Surgery: A Systematic Review. Ann. Surg. 275, e75–e81. https://doi.org/10.1097/SLA.0000000000004078
Moher, D., Schulz, K.F., Simera, I., Altman, D.G., 2010. Guidance for developers of health research reporting guidelines. PLoS Med. 7, e1000217. https://doi.org/10.1371/journal.pmed.1000217
Moons, K.G., Altman, D.G., Reitsma, J.B., Ioannidis, J.P., Macaskill, P., Steyerberg, E.W., Vickers, A.J., Ransohoff, D.F., Collins, G.S., 2015. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. Ann. Intern. Med. 162, W1–W73.
Moons, K.G.M., de Groot, J.A.H., Bouwmeester, W., Vergouwe, Y., Mallett, S., Altman, D.G., Reitsma, J.B., Collins, G.S., 2014. Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the CHARMS checklist. PLoS Med. 11, e1001744. https://doi.org/10.1371/journal.pmed.1001744

Moreno, L., Krishnan, J.A., Duran, P., Ferrero, F., 2006. Development and validation of a clinical prediction rule to distinguish bacterial from viral pneumonia in children. Pediatr. Pulmonol. 41, 331–337. https://doi.org/10.1002/ppul.20364

Morgan, J., 2011. The influence of selected demographic characteristics on the reading ability of fourth grade students in Louisiana. LSU Dr. Diss. https://doi.org/10.31390/gradschool_dissertations.2118

Muinga, N., Paton, C., Gicheha, E., Omoke, S., Abejirinde, I.-O.O., Benova, L., English, M., Zweekhorst, M., 2021. Using a human-centred design approach to develop a comprehensive newborn monitoring chart for inpatient care in Kenya. BMC Health Serv. Res. 21, 1010. https://doi.org/10.1186/s12913-021-07030-x

Munzner, T., 2009. A nested model for visualization design and validation. IEEE Trans. Vis. Comput. Graph. 15.

Murphy, C.A., Laforet, K., Da Rosa, P., Tabamo, F., Woodbury, M.G., 2012. Reliability and Predictive Validity of Inlow's 60-Second Diabetic Foot Screen Tool. Adv. Skin Wound Care 25, 261–266. https://doi.org/10.1097/01.ASW.0000415343.45178.91

Nahhas, R.W., n.d. 5.9 Interactions | Introduction to Regression Methods for Public Health Using R.

Najafabadi, A.H.Z., Ramspek, C.L., Dekker, F.W., Heus, P., Hooft, L., Moons, K.G.M., Peul, W.C., Collins, G.S., Steyerberg, E.W., Diepen, M. van, 2020. TRIPOD statement: a preliminary pre-post analysis of reporting and methods of prediction models. BMJ Open 10, e041537. https://doi.org/10.1136/bmjopen-2020-041537

National Academies of Sciences, E., Affairs, P. and G., Committee on Science, E., Information, B. on R.D. and, Sciences, D. on E. and P., Statistics, C. on A. and T., Analytics, B. on M.S. and, Studies, D. on E. and L., Board, N. and R.S., Education, D. of B. and S.S. and, Statistics, C. on N., Board on Behavioral, C., Science, C. on R. and R. in, 2019. Understanding Reproducibility and Replicability, in: Reproducibility and Replicability in Science. National Academies Press (US).

Norris, E., Hastings, J., Marques, M.M., Mutlu, A.N.F., Zink, S., Michie, S., 2021. Why and how to engage expert stakeholders in ontology development: insights from social and behavioural sciences. J. Biomed. Semant. 12, 4. https://doi.org/10.1186/s13326-021-00240-6

Norrish, G., Protonotarios, A., Stec, M., Boleti, O., Field, E., Cervi, E., Elliott, P., Kaski, J., 2023. Performance of the PRIMaCY sudden death risk prediction model for childhood hypertrophic cardiomyopathy: implications for implantable cardioverter-defibrillator decision-making. Europace 25. https://doi.org/10.1093/europace/euad330

Nuñez, S., Hexdall, A., Aguirre-Jaime, A., 2006. Unscheduled returns to the emergency department: an outcome of medical errors? Qual. Saf. Health Care 15, 102–108. https://doi.org/10.1136/qshc.2005.016618

Obermeyer, Z., Powers, B., Vogeli, C., Mullainathan, S., 2019. Dissecting racial bias in an algorithm used to manage the health of populations. Science 366, 447–453. https://doi.org/10.1126/science.aax2342

O'Connor, C., Joffe, H., 2020. Intercoder Reliability in Qualitative Research: Debates and Practical Guidelines. Int. J. Qual. Methods 19, 1609406919899220. https://doi.org/10.1177/1609406919899220

O'Donnell, A., Kaner, E., Shaw, C., Haighton, C., 2018. Primary care physicians' attitudes to the adoption of electronic medical records: a systematic review and evidence synthesis using

the clinical adoption framework. BMC Med. Inform. Decis. Mak. 18, 101. https://doi.org/10.1186/s12911-018-0703-x

Ozcan, M., Peker, S., 2023. A classification and regression tree algorithm for heart disease modeling and prediction. Healthc. Anal. 3, 100130. https://doi.org/10.1016/j.health.2022.100130

Pennello, G., Pantoja-Galicia, N., Evans, S., 2016. Comparing Diagnostic Tests on Benefit-Risk. J. Biopharm. Stat. 26, 1083–1097. https://doi.org/10.1080/10543406.2016.1226335

Phung, M.T., Tin Tin, S., Elwood, J.M., 2019. Prognostic models for breast cancer: a systematic review. BMC Cancer 19, 230. https://doi.org/10.1186/s12885-019-5442-6

Plüddemann, A., Wallace, E., Bankhead, C., Keogh, C., Van der Windt, D., Lasserson, D., Galvin, R., Moschetti, I., Kearley, K., O'Brien, K., Sanders, S., Mallett, S., Malanda, U., Thompson, M., Fahey, T., Stevens, R., 2014. Clinical prediction rules in practice: review of clinical guidelines and survey of GPs. Br. J. Gen. Pract. J. R. Coll. Gen. Pract. 64, e233-242. https://doi.org/10.3399/bjgp14X677860

Porter, C., Donthu, N., 2006. Using the technology acceptance model to explain how attitudes determine Internet usage: The role of perceived access barriers and demographics. J Bus Res.

PubMed [WWW Document], n.d. . PubMed. URL https://pubmed.ncbi.nlm.nih.gov/ (accessed 9.28.21).

REDCap [WWW Document], n.d. URL https://redcap.iths.org/ (accessed 9.22.23).

Reeder, B., Revere, D., Olson, D.R., Lober, W.B., 2011. Perceived usefulness of a distributed community-based syndromic surveillance system: a pilot qualitative evaluation study. BMC Res. Notes 4, 187. https://doi.org/10.1186/1756-0500-4-187

Richardson, S., Khan, S., McCullagh, L., Kline, M., Mann, D., McGinn, T., 2015. Healthcare provider perceptions of clinical prediction rules. BMJ Open 5. https://doi.org/10.1136/bmjopen-2015-008461

Roalfe, A.K., Mant, J., Doust, J.A., Barton, P., Cowie, M.R., Glasziou, P., Mant, D., McManus, R.J., Holder, R., Deeks, J.J., Doughty, R.N., Hoes, A.W., Fletcher, K., Hobbs, F.D.R., 2012. Development and initial validation of a simple clinical decision tool to predict the presence of heart failure in primary care: the MICE (Male, Infarction, Crepitations, Edema) rule. Eur. J. Heart Fail. 14, 1000–1008. https://doi.org/10.1093/eurjhf/hfs089

Schonberg, M.A., Li, V.W., Eliassen, A.H., Davis, R.B., LaCroix, A.Z., McCarthy, E.P., Rosner, B.A., Chlebowski, R.T., Rohan, T.E., Hankinson, S.E., Marcantonio, E.R., Ngo, L.H., 2015. Performance of the Breast Cancer Risk Assessment Tool Among Women Aged 75 Years and Older. JNCI J. Natl. Cancer Inst. 108, djv348. https://doi.org/10.1093/jnci/djv348

Scimago Journal & Country Rank [WWW Document], n.d. URL https://www.scimagojr.com/ (accessed 1.28.24).

Sezgin, E., 2023. Artificial intelligence in healthcare: Complementing, not replacing, doctors and healthcare providers. Digit. Health 9, 20552076231186520. https://doi.org/10.1177/20552076231186520

Shipe, M.E., Deppen, S.A., Farjah, F., Grogan, E.L., 2019. Developing prediction models for clinical use using logistic regression: an overview. J. Thorac. Dis. 11, S574–S584. https://doi.org/10.21037/jtd.2019.01.25

Shortliffe, E.H., Cimino, J.J. (Eds.), 2014. Biomedical Informatics: Computer Applications in Health Care and Biomedicine, 4th ed. Springer-Verlag, London.

Shrier, L.A., Burke, P.J., Jonestrask, C., Katz-Wise, S.L., 2020. Applying systems thinking and human-centered design to development of intervention implementation strategies: An example from adolescent health research. J. Public Health Res. 9. https://doi.org/10.4081/jphr.2020.1746

Simera, I., Moher, D., Hirst, A., Hoey, J., Schulz, K.F., Altman, D.G., 2010. Transparent and accurate reporting increases reliability, utility, and impact of your research: reporting guidelines and the EQUATOR Network. BMC Med. 8, 24. https://doi.org/10.1186/1741-7015-8-24

Singh, H., Giardina, T.D., Meyer, A.N.D., Forjuoh, S.N., Reis, M.D., Thomas, E.J., 2013. Types and Origins of Diagnostic Errors in Primary Care Settings. JAMA Intern. Med. 173, 418–425. https://doi.org/10.1001/jamainternmed.2013.2777

Singh, H., Meyer, A.N.D., Thomas, E.J., 2014. The frequency of diagnostic errors in outpatient care: estimations from three large observational studies involving US adult populations. BMJ Qual. Saf. 23, 727–731. https://doi.org/10.1136/bmjqs-2013-002627

Solomon, J., Dauber-Decker, K., Richardson, S., Levy, S., Khan, S., Coleman, B., Persaud, R., Chelico, J., King, D., Spyropoulos, A., McGinn, T., 2023. Integrating Clinical Decision Support Into Electronic Health Record Systems Using a Novel Platform (EvidencePoint): Developmental Study. JMIR Form. Res. 7, e44065. https://doi.org/10.2196/44065

Sperandei, S., 2014. Understanding logistic regression analysis. Biochem. Medica 24, 12. https://doi.org/10.11613/BM.2014.003

Spitzer, R.L., Kroenke, K., Williams, J.B., 1999. Validation and utility of a self-report version of PRIME-MD: the PHQ primary care study. Primary Care Evaluation of Mental Disorders. Patient Health Questionnaire. JAMA 282, 1737–1744. https://doi.org/10.1001/jama.282.18.1737

Steed, C.A., Shipman, G., Thornton, P., Ricciuto, D., Erickson, D., Branstetter, M., 2012. Practical Application of Parallel Coordinates for Climate Model Analysis. Procedia Comput. Sci., Proceedings of the International Conference on Computational Science, ICCS 2012 9, 877–886. https://doi.org/10.1016/j.procs.2012.04.094

Steyerberg, E.W., Vergouwe, Y., 2014. Towards better clinical prediction models: seven steps for development and an ABCD for validation. Eur. Heart J. 35, 1925–1931. https://doi.org/10.1093/eurheartj/ehu207

Strandberg, R., Jepsen, P., Hagström, H., 2024. Developing and validating clinical prediction models in hepatology – An overview for clinicians. J. Hepatol. 0. https://doi.org/10.1016/j.jhep.2024.03.030

Sutton, R.T., Pincock, D., Baumgart, D.C., Sadowski, D.C., Fedorak, R.N., Kroeker, K.I., 2020. An overview of clinical decision support systems: benefits, risks, and strategies for success. Npj Digit. Med. 3, 1–10. https://doi.org/10.1038/s41746-020-0221-y

Thomas, B., 2011. E-mail Address Harvesting on PubMed—A Call for Responsible Handling of E-mail Addresses. Mayo Clin. Proc. 86, 362. https://doi.org/10.4065/mcp.2010.0817

United Nations, n.d. World Population Prospects - Population Division - United Nations [WWW Document]. URL https://population.un.org/wpp/DefinitionOfRegions/ (accessed 9.25.23).

Van Belle, V., Van Calster, B., 2015. Visualizing Risk Prediction Models. PLoS ONE 10. https://doi.org/10.1371/journal.pone.0132614

Van Calster, B., McLernon, D.J., van Smeden, M., Wynants, L., Steyerberg, E.W., Bossuyt, P., Collins, G.S., Macaskill, P., McLernon, D.J., Moons, K.G.M., Steyerberg, E.W.,

Van Calster, B., van Smeden, M., Vickers, A.J., On behalf of Topic Group 'Evaluating diagnostic tests and prediction models' of the STRATOS initiative, 2019. Calibration: the Achilles heel of predictive analytics. BMC Med. 17, 230. https://doi.org/10.1186/s12916-019-1466-7

van Mens, H.J.T., Duijm, R.D., Nienhuis, R., de Keizer, N.F., Cornet, R., 2020. Towards an Adoption Framework for Patient Access to Electronic Health Records: Systematic Literature Mapping Study. JMIR Med. Inform. 8, e15150. https://doi.org/10.2196/15150

Venkatesh, V., 2000. Determinants of Perceived Ease of Use: Integrating Control, Intrinsic Motivation, and Emotion into the Technology Acceptance Model. Inf. Syst. Res. 11, 342–365.

Venkatesh, V., Bala, H., 2008. Technology Acceptance Model 3 and a Research Agenda on Interventions. Decis. Sci. 39, 273–315. https://doi.org/10.1111/j.1540-5915.2008.00192.x

Verma, N., Fleischmann, K., Koltai, K., 2018. Demographic factors and trust in different news sources. Proc. Assoc. Inf. Sci. Technol. 55, 524–533. https://doi.org/10.1002/pra2.2018.14505501057

Verplanken, B., Orbell, S., 2022. Attitudes, Habits, and Behavior Change. Annu. Rev. Psychol. 73, 327–352. https://doi.org/10.1146/annurev-psych-020821-011744

Vickers, A., Van Calster, B., Steyerberg, E., 2019. A simple, step-by-step guide to interpreting decision curve analysis. Diagn. Progn. Res. 3. https://doi.org/10.1186/s41512-019-0064-7

Vickers, A.J., Holland, F., 2021. Decision curve analysis to evaluate the clinical benefit of prediction models. Spine J. Off. J. North Am. Spine Soc. 21, 1643–1648. https://doi.org/10.1016/j.spinee.2021.02.024

Waljee, A.K., Higgins, P.D.R., Singal, A.G., 2014. A Primer on Predictive Models. Clin. Transl. Gastroenterol. 5, e44. https://doi.org/10.1038/ctg.2013.19

Wallace, E., Johansen, M.E., 2018. Clinical Prediction Rules: Challenges, Barriers, and Promise. Ann. Fam. Med. 16, 390–392. https://doi.org/10.1370/afm.2303

Wallace, E., Uijen, M.J.M., Clyne, B., Zarabzadeh, A., Keogh, C., Galvin, R., Smith, S.M., Fahey, T., 2016. Impact analysis studies of clinical prediction rules relevant to primary care: a systematic review. BMJ Open 6, e009957. https://doi.org/10.1136/bmjopen-2015-009957

Walsh, B.T., Bookheim, W.W., Johnson, R.C., Tompkins, R.K., 1975. Recognition of streptococcal pharyngitis in adults. Arch. Intern. Med. 135, 1493–1497.

Walsh, C.G., McKillop, M.M., Lee, P., Harris, J.W., Simpson, C., Novak, L.L., 2021. Risky business: a scoping review for communicating results of predictive models between providers and patients. JAMIA Open 4, ooab092. https://doi.org/10.1093/jamiaopen/ooab092

Wang, R.Y., Strong, D.M., 1996. Beyond accuracy: What data quality means to data consumers. J. Manag. Inf. Syst. 12, 5–33.

Weissgerber, T.L., Milic, N.M., Winham, S.J., Garovic, V.D., 2015. Beyond Bar and Line Graphs: Time for a New Data Presentation Paradigm. PLOS Biol. 13, e1002128. https://doi.org/10.1371/journal.pbio.1002128

Whiting, P.F., Rutjes, A.W.S., Westwood, M.E., Mallett, S., Deeks, J.J., Reitsma, J.B., Leeflang, M.M.G., Sterne, J.A.C., Bossuyt, P.M.M., QUADAS-2 Group, 2011. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann. Intern. Med. 155, 529–536. https://doi.org/10.7326/0003-4819-155-8-201110180-00009

Wilson, P.W., D'Agostino, R.B., Levy, D., Belanger, A.M., Silbershatz, H., Kannel, W.B., 1998. Prediction of coronary heart disease using risk factor categories. Circulation 97, 1837–1847. https://doi.org/10.1161/01.cir.97.18.1837

Wolff, R.F., Moons, K.G.M., Riley, R.D., Whiting, P.F., Westwood, M., Collins, G.S., Reitsma, J.B., Kleijnen, J., Mallett, S., 2019. PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies. Ann. Intern. Med. 170, 51–58. https://doi.org/10.7326/M18-1376

Yang, C., Kors, J.A., Ioannou, S., John, L.H., Markus, A.F., Rekkas, A., de Ridder, M.A.J., Seinen, T.M., Williams, R.D., Rijnbeek, P.R., 2022. Trends in the conduct and reporting of clinical prediction model development and validation: a systematic review. J. Am. Med. Inform. Assoc. JAMIA 29, 983–989. https://doi.org/10.1093/jamia/ocac002

Yin, J., Zhu, Q., Zhang, K., Gao, W., Wu, J., Lu, Z., Jiang, K., Miao, Y., 2022. Development and validation of risk prediction nomogram for pancreatic fistula and risk-stratified strategy for drainage management after pancreaticoduodenectomy. Gland Surg. 11, 42–55. https://doi.org/10.21037/gs-21-550

Yusuf, M., Atal, I., Li, J., Smith, P., Ravaud, P., Fergie, M., Callaghan, M., Selfe, J., 2020. Reporting quality of studies using machine learning models for medical diagnosis: a systematic review. BMJ Open 10, e034568. https://doi.org/10.1136/bmjopen-2019-034568

Zigman Suchsland, M.L., Rahmatullah, I., Lutz, B., Lyon, V., Huang, S., Kline, E., Graham, C., Cooper, S., Su, P., Smedinghoff, S., Chu, H.Y., Sewalk, K., Brownstein, J.S., Thompson, M.J., on behalf of Seattle Flu Study investigators, 2021. Evaluating an app-guided self-test for influenza: lessons learned for improving the feasibility of study designs to evaluate self-tests for respiratory viruses. BMC Infect. Dis. 21, 617. https://doi.org/10.1186/s12879-021-06314-1

Zuithoff, N.P., Vergouwe, Y., King, M., Nazareth, I., Hak, E., Moons, K.G., Geerlings, M.I., 2009. A clinical prediction rule for detecting major depressive disorder in primary care: the PREDICT-NL study. Fam. Pract. 26, 241–250. https://doi.org/10.1093/fampra/cmp036

APPENDIX A

No	Question
1	What type of provider are you? **Options:** MD PA DO ARNP MD Resident
2	How many years have you been in practice in primary care since the end of your professional training? **Options:** ≤ 5 years > 5 - ≤ 10 years > 10 - ≤ 15 years > 15 - ≤ 20 years > 20 years
3	How much experience do you have using clinical prediction rules to either diagnose a disease or determine prognosis of a disease during patient encounters in primary care? **Options:** Significant experience Moderate experience Little experience No experience
4	Select the three factors that motivate you most to use clinical prediction rules during patient encounters in primary care. (Select up to three) **Options:** • Easy to use • Improve diagnostic/prognostic accuracy • Avoid unnecessary cost • Improve the communication of diagnosis/ prognosis wit patients • Save time to diagnose/ prognose • None of the above • I don't have an opinion because I have no experience in using CPRs
5	Select the top three clinical prediction rules that you believe are most useful in primary care. (Select up to three) **Options:** **Flu Score**: Support seasonal flu diagnosis using influenza-like illness, such as cough and fever **Walsh Rule/ Centor Score:** Distinguish patients with viral infections and those suspected of having streptococcal pharyngitis **Bacterial Pneumonia Score**: Identify children with pneumonia who need antibiotic medication

No	Question
	Framingham Score/ QRISK 2: Provide individualized cardiovascular disease risks to high-risk patients **MICE Rule**: Optimize referral to echocardiography for patients with suspected heart failure **GERD Score**: Support the diagnosis of Gastro-Esophageal Reflux Disease (GERD) for patients with upper gastrointestinal complaints **PHQ-9/ PREDICT-NL**: Detect patients with major depressive disorder. **The Breast Cancer Risk Assessment Tool (BCRAT)/ Gail model/ Nottingham Prognostic Index**: Predict breast cancer risk and prognosis **Risk Assessment Tool (RAT)/ QCancer**: Estimate risk for patients with symptoms of possible cancer **Diabetic Foot Screen Tool**: identify patietns at risk for diabetic foot ulcers **predictAL:** Predict the onset of hazardous alcohol drinking over 12 months **Marburg Heart Score**: Rules out coronary artery disease in patients with chest pain **CHADS2:** Classification schemes that estimate stroke risk in patients with Atrial Fibrillation. **"CURB" severity score**: stratify patients with Community Acquired Pneumonia CAP into different management groups using six-point score based on confusion, urea, respiratory rate, blood pressure, and age.
6	Which of the following are barriers that you experience to using clinical prediction rules in primary care? (Select up to three) **Options:** • Disrupt clinical workflow • Difficult to communicate the results with patients • Difficult to interpret • Do not improve accuracy • Interference with clinicians' autonomy • Time-consuming • None of the above • I don't have an opinion because I have no experience in using CPRs
7	To what extent do you agree that your use of clinical prediction rules is well supported by your electronic health record (EHR)? **Options:** Strongly disagree Disagree Neither agree nor agree Agree Strongly agree
8	To what extent do you agree that clinical prediction rules **helps you to improve your diagnostic and prognostic accuracy** during patient encounters in primary care?

No	Question
	Options: Strongly disagree Disagree Neither agree nor agree Agree Strongly agree
9	To what extent do you agree that clinical prediction rules are **easy for you to use** during patient encounters in primary care? **Options:** Strongly disagree Disagree Neither agree nor agree Agree Strongly agree
10	To what extent do you agree that clinical prediction rules **help you to communicate disease diagnosis or prognosis** during patient encounter in primary care? **Options:** Strongly disagree Disagree Neither agree nor agree Agree Strongly agree

APPENDIX B

PubMed search strategy for identifying articles on clinical prediction models in primary care

The following search strategy was employed to identify relevant articles related to clinical prediction models aimed for use in primary care settings. The search was limited to articles published between January 1, 2015, and October 31, 2022.

Prediction Model Keywords (#1)
"prediction model*"[Title/Abstract] OR "prediction rul*"[Title/Abstract] OR "predictive model*"[Title/Abstract] OR "predictive rul*"[Title/Abstract] OR "predictive scor*"[Title/Abstract] OR "prediction scor*"[Title/Abstract] OR "clinical model"[Title/Abstract] OR "decision rule"[Title/Abstract] OR "Predictive Value of Tests"[MeSH Terms] OR "Machine Learning"[MeSH Terms] OR "Clinical Decision Rules"[MeSH Terms] OR "Probability Learning"[MeSH Terms] OR "Bayes Theorem"[MeSH Terms] OR "Forecasting"[MeSH Terms] OR "prognostic model" OR "diagnostic model"

Prognosis and Diagnosis Keywords (#2)
"Diagnosis"[MeSH Terms] OR "diagnos*"[Title/Abstract] "Prognosis"[MeSH Terms] OR "prognos*"[Title/Abstract]

Disease-Specific Keywords Related to Primary Care (#3)
“type 2 Diabetes”[Title/Abstract] OR “Diabetes type 2”[Title/Abstract] OR “Diabetes mellitus”[Title/Abstract] OR "Diabetes Mellitus"[Mesh] OR "Hypertension"[Mesh] OR hypertension[Title/Abstract] OR "Respiratory Tract Infections"[Mesh] OR "Depression"[Mesh] OR "Depressive Disorder"[Mesh] OR depression[Title/Abstract] OR "Anxiety"[Mesh] OR "Anxiety"[Title/Abstract] OR "Back Pain"[Mesh] OR "Low Back Pain"[Mesh] OR "Back Pain"[Title/Abstract] OR "Arthritis"[Mesh] OR "Arthritis"[Title/Abstract] OR "Dermatitis"[Mesh] OR "Dermatitis"[Title/Abstract] OR "Otitis Media"[Mesh] OR "Otitis Media"[Title/Abstract] OR "Pneumonia"[Mesh] OR "Pneumonia"[Title/Abstract] OR "Tuberculosis"[Mesh] OR "Tuberculosis"[Title/Abstract] OR "Parasites"[Mesh] OR "Parasites"[Title/Abstract] OR "Anemia"[Mesh] OR "Anemia"[Title/Abstract] OR "HIV"[Mesh] OR "HIV"[Title/Abstract] OR "Cough"[Mesh] OR "Cough"[Title/Abstract] OR "Pharyngitis"[Mesh] OR "Pharyngitis"[Title/Abstract] OR "Fever"[Mesh] OR "Fever"[Title/Abstract] OR "Headache"[Mesh] OR "Headache"[Title/Abstract] OR "Fatigue"[Mesh] OR "Fatigue"[Title/Abstract] OR "Sinusitis"[Mesh] OR "Sinusitis"[Title/Abstract] OR "Cardiovascular Diseases"[Mesh] OR "Cardiovascular Disease"[Title/Abstract] OR "Vertigo"[Mesh] OR "Vertigo"[Title/Abstract] OR "Dizziness"[Mesh] OR "Dizziness"[Title/Abstract] OR "Asthma"[Mesh] OR "Asthma"[Title/Abstract] OR "Tonsillitis"[Mesh] OR "Tonsillitis"[Title/Abstract] OR "Dyspepsia"[Mesh] OR "Dyspepsia"[Title/Abstract] OR "Bronchitis"[Mesh] OR "Bronchitis"[Title/Abstract] OR "Bronchiolitis"[Mesh] OR "Bronchiolitis"[Title/Abstract] OR "Epilepsy"[Mesh] OR "Epilepsy"[Title/Abstract]

Time Range (#4)

("2015/01/01"[PDAT] : "2022/10/31"[PDAT])

Final Combined Query
#1 AND #2 AND #3 AND #4

APPENDIX C

ABOUT YOUR EXPERIENCE AS AUTHOR

The following questions ask about your experience as a author of prediction model research papers in **peer-reviewed biomedical journals or peer-reviewed conference proceedings**.

1. How many prediction model research papers have you published as a **author**?
   - 0
   - 1 – 5
   - 6 – 10
   - 11 – 20
   - more than 20
2. How many years have you been involved in publishing prediction model research papers as **a author**?
   - Less than a year
   - 1 to 5 years
   - 6 to 10 years
   - 11 to 20 years
   - More than 20 years
3. What kind of prediction model research paper(s) have you published as **a author**? (Please select all that apply)
   - Developing prediction models for **diagnosis**
   - Validating prediction models for **diagnosis**
   - Developing prediction models for **prognosis**
   - Validating prediction models for **prognosis**

- Other types of prediction model research papers (e.g., developing and validating prediction models for non-clinical purposes, systematic reviews, meta-analysis, opinion)

please describe: (fill in)

ABOUT YOUR EXPERIENCE AS A AUTHOR TO ENSURE UNDERSTANDABLE, USEFUL, AND TRUSTWORTHY STUDY RESULTS

The following questions ask you as a author how difficult it is to ensure that the study results of prediction model research papers are understandable, useful, and trustworthy in **peer-reviewed biomedical journals or peer-reviewed conference proceedings**.

This survey uses the following definitions for **understandable**, **useful,** and **trustworthy study results.**

- **U*nderstandable* study results** refer to the degree to which a user is able to interpret, exemplify, classify, summarize, infer, compare, and explain the study results of prediction model research papers.
- **U*seful* study results** refer to the degree to which a user is satisfied with their perceived achievement of pragmatic goals, including the goals after reading the study results of prediction model research papers.
- **T*rustworthy* study results** refer to the degree to which a user has confidence that the study results of prediction model research papers provide information as they should.

4. As a author, how difficult is it to author **understandable** study results in prediction model research papers?
   - Verry difficult
   - Difficult
   - Neutral

- Easy
- Very easy
- I have never authored prediction model research papers.

5. As a author, how difficult to author **useful** study results in prediction model research papers?
   - Verry difficult
   - Difficult
   - Neutral
   - Easy
   - Very easy
   - I have never authored prediction model research papers.
6. As a author, how difficult is it to author **trustworthy** study results in prediction model research papers?
   - Verry difficult
   - Difficult
   - Neutral
   - Easy
   - Very easy
   - I have never authored prediction model research papers.
7. Would you comment on why on your answers to questions 4 - 6?

ABOUT YOUR EXPERIENCE AS A PEER REVIEWER

The following questions ask about your experience as a peer reviewer of prediction model research papers in **peer-reviewed biomedical journals or peer-reviewed conference proceedings**.

8. How many prediction model research papers have you served as **a peer reviewer**?
   - 0
   - 1 – 5
   - 6 – 10
   - 11 – 20
   - more than 20
9. How many years have you been involved in publishing prediction model research papers as **a peer reviewer**?
   - Less than a year
   - 1 to 5 years
   - 6 to 10 years
   - 11 to 20 years
   - More than 20 years
10. What kind of prediction model research paper(s) have you served as **a peer reviewer**? (Please select all that apply)
    - Developing prediction models <u>for diagnosis</u>
    - Validating prediction models <u>for diagnosis</u>
    - Developing prediction models <u>for prognosis</u>
    - Validating prediction models <u>for prognosis</u>
    - Other types of prediction model research papers (e.g., developing and validating prediction models for non-clinical purposes, systematic reviews, meta-analysis, opinion) please describe: (fill in)

ABOUT YOUR EXPERIENCE AS A PEER REVIEWER TO ENSURE UNDERSTANDABLE, USEFUL, AND TRUSTWORTHY STUDY RESULTS

The following questions ask you as a peer reviewer how difficult it is to provide feedback to authors to ensure that the study results of prediction model research papers are understandable, useful, and trustworthy in **peer-reviewed biomedical journals or peer-reviewed conference proceedings**.

This survey uses the following definitions for **understandable**, **useful,** and **trustworthy study results.**

- **U*nderstandable* study results** refer to the degree to which a user is able to interpret, exemplify, classify, summarize, infer, compare, and explain the study results of prediction model research papers.
- **U*seful* study results** refer to the degree to which a user is satisfied with their perceived achievement of pragmatic goals, including the goals after reading the study results of prediction model research papers.
- **T*rustworthy* study results** refer to the degree to which a user has confidence that the study results of prediction model research papers provide information as they should.

11. As a reviewer, how difficult is it to provide feedback to authors to ensure **understandable study results** in prediction model research papers?
    - Verry difficult
    - Difficult
    - Neutral
    - Easy
    - Very easy

- I have never reviewed prediction model research papers.

12. As a reviewer, how difficult is it to provide feedback to authors to ensure **useful study results** in prediction model research papers?
    - Verry difficult
    - Difficult
    - Neutral
    - Easy
    - Very easy
    - I have never reviewed prediction model research papers.
13. As a reviewer, how difficult is it to provide feedback to authors to ensure **trustworthy study results** in prediction model research papers?
    - Verry difficult
    - Difficult
    - Neutral
    - Easy
    - Very easy
    - I have never reviewed prediction model research papers.
14. Would you comment on why on your answers to questions 11 - 13?

ABOUT FINDING UNDERSTANDABLE, USEFUL, AND TRUSTWORTHY STUDY RESULTS

The following questions ask how often you find the study results in prediction model research papers in peer-reviewed biomedical journals or peer-reviewed conference proceedings to be understandable, useful, and trustworthy.

This survey uses the following definitions for **understandable**, **useful,** and **trustworthy study results.**

- **U*nderstandable* study results** refer to the degree to which a user is able to interpret, exemplify, classify, summarize, infer, compare, and explain the study results of prediction model research papers.
- **U*seful* study results** refer to the degree to which a user is satisfied with their perceived achievement of pragmatic goals, including the goals after reading the study results of prediction model research papers.
- **T*rustworthy* study results** refer to the degree to which a user has confidence that the study results of prediction model research papers provide information as they should.

15. How often do you find the study results in prediction model research papers **understandable**?
    - Never
    - Rarely
    - Sometimes
    - Often
    - Always

16. How often do you find the study results in prediction model research papers **useful**?
    - Never
    - Rarely
    - Sometimes
    - Often

- Always

17. How often do you find the study results in prediction model research papers **trustworthy**?

- Never
- Rarely
- Sometimes
- Often
- Always

ABOUT OTHER QUALITY CHARACTERISTICS

This study focuses on understandable, useful, and trustworthy study results as quality characteristics in prediction model research papers.

18. Beyond understandable, useful, and trustworthy study results of prediction model research papers, what other characteristics do you think are important? (Fill in)

ABOUT YOU

The following questions ask about your demographic information.

19. In what type of organization(s) do you currently work? (Please select all that apply)

- Academia
- Industry
- Government
- Consultancy
- Freelancing
- Other, please describe: (fill in)
- Prefer not to say

20. Which age category are you in?

- 18 – 24 years old
- 25 – 34 years old
- 35 – 44 years old
- 45 – 65 years old
- 66 – 79 years old
- 80 years old and over
- Prefer not to say

21. Which gender do you identify with?

- Man
- Woman
- I prefer to describe: (fill in)
- Prefer not to say

22. Which country are you from?

Would you like to discuss your answers in more detail through an optional 1-hour remote interview?

Participants who complete the optional interview will receive a $25 Tango gift card after the interview.

__Yes

Please enter your email to schedule your interview:

__No

Thank you for your participation!

APPENDIX D

ID	Text	Reasons for exclusion
5	you can spin data both ways	incomplete ideas
14	It depends!	
17	My study had low numbers and proposed a possible predictive model based on our data, but couldn't validate because of low numbers.	outside the scope of challenges in authoring or reviewing to ensure understandable, useful, and trustworthy results in CPM studies
43	Antwerpen depends on type of study and prediction model	incomplete ideas
52	Since this type of research was my first, I was challenged alittle bit.... but result were good..	outside the scope of challenges in authoring or reviewing to ensure understandable, useful, and trustworthy results in CPM studies
62	We're dependent on skilled and trustworthy biostatisticians.	incomplete ideas
82	This is relative, depending on the complexity of the subject matter and the analysis performed by the authors.	incomplete ideas
105	Worked mainly in population studies or small clinical studies, both have various limitations (heterogeneity, statistical power, chance findings).	incomplete ideas
150	4. depends on context	incomplete ideas
30	I do all my stat myself	incomplete ideas

APPENDIX E

ID	Text	Theme
209	Clinical co-authors have a desire to put a positive spin on thier prediction model - they rarely are happy to take a deep dive into what the results really mean - rather they take any C-statistic at face value and assume anything ~0.8 must mean the model works (that is not the case mostly).	Difficulties in understanding clinical prediction models among the target users
7	There are statistical subtleties in what makes a good clinical prediction model. For the most part these are explainable though certain concepts are likely to be quite difficult for a general audience. So I picked neutral as in between.	
9	I think sometimes the issue is between the level of statistical detail that a prediction model should have in a paper, and the amount of statistical detail that an average clinical reader would be interested in. Though much of this could go in a supplement.	
37	Many people (including physicians and patients) misinterpret risk prediction models as causal.	
63	understandability is difficult and depends on the knowledge of the audience which is out of one's control;	
93	Medical prediction is extremely challenging in all three respects--developing and validating models is hard, both because of the complexity of the underlying biology and the limitations of validation data. Because the models are complicated, explaining them is challenging too.	
94	I don't know that any of these are easy or difficult because the degree to which a user can utilize the paper depends on the training of the user, over which I have no control.	
107	The barrier to being useful for a clinician is extremely high because they are skeptical and stuck in their ways. At the same time, inadequate training makes it hard for the target audience to understand the outputs.	
117	Study results can be understandable to those knowledgeable about the study topic, but less so to others.	

ID	Text	Theme
127	Publishing understandable results depends on the reader's knowledge and expertise.	
143	Making a good prediction model is difficult. Getting readers to understand that a prediction model is a tool that should be taken in context with other factors is near impossible.	
198	It is a challenge to describe statistical analyses and results of a prediction model paper in understandable language for the target audience, which typically includes health care providers. Understanding the results and implications of prediction models does require specific statistical and methodological knowledge.	
229	AI models are almost by definition more challenging to be understandable - I have not been involved in conducting them but have written with others on the methodological/conceptual issues	
231	I don't think it is super hard to publish about this topic. I do think that many people do not understand model, prediction, and model developments.	
236	I think these questions hit upon the difficulty of conveying the limitations of a study as well as the real world impact of such a study, particularly to a clinical audience.	
239	Hard to achieve accurate models for multifactorial risks.	
240	Very technical matter, which is often hard to translate to clinical audiences	
260	It is always difficult for me to make doctors understand the results of my studies.	
68	machine learning based prediction models are non interpretable in terms of directionality of results	
73	I am an expert in sick leave research, not in statistics or prediction models. The biostatisticians who developed the models sometimes worked beyond my understanding of statistics, so while I could have opinions on the predictors that were included, I sometimes struggled to interpret and evaluate the model output that they presented.	
75	It is easy to tell authors how it should be done. It is harder for them to do it.	

ID	Text	Theme
8	Some metrics, e.g. for calibration and internal validation which are necessary for developing and reporting prognostic risk models are difficult to communicate in a way which is easily understood by people less statistical.	
22	But point 4 is tricky because often there is no link to decision making. Without that link it is hard to convince clinicians to use prediction and they are used rarely indeed	Difficulties in presenting clinically relevant and useful prediction models
41	Well, it is one thing to build a predictive model that's significant using a cutting edge methodology. But it is a totally different ballgame if one wishes for these results to be useful, meaningful, and reliable to clinicians.	
79	The art of writing prediction models is to be sure that they are clinically useful and actionable (you would treat low and high-risk patients differently) and if those are met, then it is not difficult to explain.	
84	In my experience it is much easier to develop a prediction model than it is to developed a useful prediction model. And one has to take extra care to develop a useful and trustworthy model.	
87	If the definition of "useful" is to provide information about the prediciton model performande - then it is not super hard. If the definition of useful is to provide models that could improve clinical outcomes among patients, then it is very very hard.	
111	The key challenge is delivering useful results, because ultimately the key useful result would be randomized trial. But that is very expensive and time consuming. So short of that, how do you prove utility? Very difficult... Being clear and trustworthy isn't that difficult	
127	Many prediction models are published but they are rarely useful clinically. This may be because of the irrelevance of the question or the limitations of the study (sample size, model prediction etc.). Furthermore, models are not useful unless they're validated in different cohorts.	
127	Many prediction models are published but they are rarely useful clinically. This may be because of the irrelevance of the question or the limitations of the study (sample size, model prediction etc.). Furthermore, models are not useful unless they're validated in different cohorts.	

ID	Text	Theme
142	Many diagnostic and prognostic manuscripts do not focus on the early recognition of clinically important outcomes - which reduces their utility and impact.	
148	Useful require additional studies to development or validation as they require assessment in comparison to standard practice and intervention studies. This increases the difficulty.	
148	However, it is much more difficult to talk of usefulness as manuscripts rarely compare against standard practice.	
161	I'm not sure I understand how you are defining and using the above terms. I'm a qualitative researcher and I see a disconnect between qual and quant researchers. My main focus is on how usable and understandable the models are for providers and patients. The quant researchers I work with seem to believe that their models are solely logic- and evidence-based, and are not always sympathetic to providers' (who may not be as mathematically savvy as the quant researchers) suspicions and inability to implement their model. They seem particularly uninterested in political reasons for not using the models properly.	
210	I think it is easy to understand the potential usefulness of the models but have not seen many implemented in practice. Is this because the models are too difficult to adopt, or are the models not robust enough to reproduce? etc.	
220	Useful study results can be hard sometimes when there are null findings related to the predictive model or the study design was underpowered to account for all the variables we wanted to assess.	
221	in my experience many primary authors of risk prediction papers are methodologists and non-clinical (e.g. PhD) researchers. The biggest weakness I see in prediction models are that reviewers oversell clinical importance /utility of the model or develop clinical prediction models that don't add a great deal of additional benefit from simpler risk calculation strategies	
227	I did not have much experience but I often did not feel that results were very useable in clinical practice	
249	All analysis involving prediction models are hard and need an expert team of researchers to conduct it in a useful, understandable and trustworthy way.	

ID	Text	Theme
254	Translating models requires buy-in from diverse group of experts- it is quite challenging to achieve the above.	
66	As mentioned in my prior response, I think there are plenty of risk algorithms being developed, but I'm just not sure how many are useful or implementable into daily clinical practice. So it requires some nice words in the review about how the findings may be interesting but not always sure they are practical, believable, easy to use...	
74	Exactly what we mean by useful is a little difficult to determine. Sometimes a model might get misleading results or never have a chance of entering practice, but still be useful because it advances the field's approach to models.	
82	Many papers end with an evaluation/validation of the predictive value of a model, so the discussion on how exactly to use model predictions' in care is usually not really touch upon and as such also does not often discussed.	
82	The use of prediction models refers to when take clinical action based on prediction models' results. This is rarely explicitly defined/evaluated, and involves difficult discussions	
10	Often difficult to predict ways in which prediction model will be clinically useful. Final models sometimes perform well but are not appropriate for day to day clinical impact because of barriers to use/data required.	
26	It is challenging to turn complex models into something simple and practical for clinicians.	
24	Prediction of outcomes is affected by a large number of factors, being able to design and carry out a study that accurately addresses these and is seen as valid in the eyes of the public is difficult.	Difficulties in ensuring trustworthy clinical prediction models
27	needs to have a practical idea, a good database with most required variables. a biostatistician, expertise on the topic	
29	It is difficult to explain to authors that their own concepts are commonly not a) properly validated, b) replicated in any way, or c) understandable to me (=others)	
73	If the results are based on open data and reproducible codes with a good consideration of know confounders and if I can see the analysis pipelines, then the result is at least trustworthy no matter what.	

ID	Text	Theme
87	As in previous comment - depends on what useful means in this context. Some of the authors have themselves (seemingly) not understood what a prediction model is and how to interpret performance metrics.	
107	One of the key challenges is trustworthiness because of poor knowledge of the situations where prediction models miss, black box models, history of coding in inherent sources of bias, and inherent distrust of automation.	
107	One of the key challenges is trustworthiness because of poor knowledge of the situations where prediction models miss, black box models, history of coding in inherent sources of bias, and inherent distrust of automation.	
236	As a reviewer I find myself spending way too much time ensuring that sufficient information is present in the manuscript so that a knowledgeable reader could understand what was done. That being said, without code and a bit more transparency on the approach it can be difficult to ensure trustworthy results -- or perhaps accurate results (i.e., that no mistakes have accidentally been made).	
242	It is difficult to ensure a valid modeling and validation framework while also having the pressure for publishing.	
66	Too many risk prediction algorithms these days, so the big challenge is determining WHY a particular risk model is useful, and whether you can really apply it clinically to patients at risk (ie, trustworthy).	
75	If studies are done well it is easy (for me) to make them clinically relevant. If they are not validated, they are less trustworthy.	
88	Hard to comment on trustworthy study results as a reviewer. Tend to assume truth unless the data appear inconsistent	
15	Sometimes is difficult to trust study results due to the difficulty of the studies reporting how handled the data, such as how they handled missing data in the study.	
29	The most difficult is to make sure prediction models are trustworthy as it involves data quality checks and sound methodology checks.	
30	a lot of authors confuse useful with statistically significant. lots of paper do not use the right control population hence not trustworthy or useful	

ID	Text	Theme
33	When validating prediction models, your results rely on the trustworthiness of the prediction models - even if your own data that you use for prediction is perfectly reliable, the results from your study are affected by any possible shortcomings in the models.	
36	I'm an researcher of Kawasaki disease. The performance of prediction model in different populations is relatively poor. So many factors can affect the trustworthy study results, such as risk of bias and what not	
10	Usefulness is more challenging since it depends on the quality of the data used to derive the prediction equations. If the data are clinic-based, for example, the usefulness of the prediction equation when applied to the general population is limited.	Difficulties in acquiring and utilizing quality data
10	Acquiring the excellent data required to develop useful results is the most challenging aspect. Rubbish in, rubbish out.	
19	Difficult because it is difficult to find data, and also find the right topic to develop a relevant prediction model.	
35	Concerning the models for the prediction of intracranial aneurysm thrombosis, there is a lack of in-vitro studies and data from real cases	
39	As a clinician, developing clinical prediction models requires independent variables that are easily identifiable symptoms and signs and easily accessible and affordable lab varaibles	
63	Usefulness depends on the quality of data and methods and can be easy for an experienced investigator;	
86	In health, where data cannot generally be shared, there are problems of replicability.	
106	Due to restrictions on clinical data access and further use/dissemination, it is extremely difficult to produce useful and reliable study results. Many studies focus on a specific region or even an organization, limiting the variability of the data used and even the number of attributes available to develop the models. Furthermore, the lack of clinical standards implemented in health information systems around the world is a major constraint in the development/validation of predictive models, as it is	

ID	Text	Theme
	impossible to recreate/reproduce a study conducted in one hospital for another given the lack of syntactic and semantic interoperability.	
117	Because of study limitations, e.g., small sample size or inability to control for certain confounders, study results may be less useful than the researcher would like.	
163	Diagnostic/prognostic modeling is often limited by availability of data and opportunity to test/deploy models in practice	
165	The major problem is that data to develop a prediction model are specific to derivation cohorts of patients, with difficulties to have reliable external validation	
171	External validation is key to generating reliable prediction models. I work in the rare disease space, so this it is often challenging to assemble suitable validation datasets.	
172	Very difficult in my area to collect data on the large number of subjects required for stable and precise prediction intervals; hard to explain to colleagues and patients about why the estimates are imprecise and also why some known risk factors make their way into the final model while others don't	
179	I feel the difficulties arise from a few areas, such as: 1- Choosing a sensible clinical question and endpoint to predict. This as a starting point is essential to creating a prediction model which will be of clinical utility.	
181	There are a lot of prediction model research papers but most are not that helpful (difficult to use, limited sample set, or the endpoint isn't that useful)	
184	Getting a reliable, well calibrated, validated prediction model from clinical datasets with missing data and small sample size is fraught with problems.	
187	The main difficulty is to find solid and complete data to develop the model or even validate a model.	

ID	Text	Theme
196	4 - it is not easy to improve on the state of the art or implement new models that perform well, many times due to data constraints;	
198	Furthermore, after development a prediction model typically should not be used directly. This requires additional research including external validation, studying the impact or implementation of the model, and if necessary, updating or adjusting the model to better fit the setting in which the model is to be used. A clinical prediction model might therefore not be directly useful to the target audience (e.g. health care providers), and therefore providing 'useful' results can be difficult.	
204	Neutral to Easy - the reason or enabling factor is do you have data. Clinical data can be quite scarce at the density to accurately check prediction - this is the major limiting factor for many groups (IMO)	
204	Equally, lack of clinical data or full clinical data hinders their ability to validate.	
208	It is difficult to obtain the original data and to replicate the authors' experiments.	
259	One of the biggest issues that we ran into was trying to truly understand and represent the population in our cohort to allow for an understanding of how translatable/ generalizable our data were to other groups	
79	Predictive models are only as good as the underlying data sources and access to large, complete and accurate datasets can be difficult. In addition, the models are often able to tell us with accuracy who will be at risk for a certain outcome but not why. Thus there is often a remaining gap between identifying s person at risk and successfully applying the appropriate intervention in a timely way.	
3	inspecting the representativeness of collected data and rationality of applied model as well as the validation methods and results	
33	Because we understand the model and its limitation, especially we cannot get all possible data for machine learning.	
11	A lot of integrity of the data is assumed. and expertise are needed. For example I reviewed paper on cardiac comorbidities that did not include diabetes because the data was not collected. Now this would render the study perfectly useless in my opinion as it is such a predictive factor	

ID	Text	Theme
11	4- in the field of transplant, a lot of the information available for risk prediction models are only available moments before transplantation or post-transplantation which begs the question of its value in preventing measures 5) different times and methods are used in different units and countries making data interpretation precarious 6) as above, but applicability of data differs across the continent	
15	Useful study results are difficult to achieve due to the difficulty to preprocess the data and interpret it in a correct way which sometimes leads to incomprehensible study results.	
20	Models are often developed using retrospective data, so prospective application can then be questionable. Often use ICD codes for diagnoses, while humans use clinical judgment and other data.	
35	In most cases, the perfect data is not available.	
58	"Difficult" is obviously a subjective term, and one sensitive to respondent's baseline-for-comparison and what aspects of difficulty they are focused on. My response is based on the fact that others make predictive claims based on within-sample predictions (i.e., without a validation/test sample) or other less-than-best practices, I think that rigorous prediction actually is more difficult than what is commonly seen in the literature. But it is not more difficult, say, than a case-control mechanistic paper.	Difficulties in navigating complexity and evolving methodologies conducting studies in clinical prediction models
63	trustworthiness can be difficult with black box approaches, even simple things like selection procedures.	
95	Many variables, possible correlations, difficulty in adjusting for confounders	
118	The methodology is straightforward and rather easy to describe. However it is difficult to have appropriate samples to develop and to validate a clinical prediction model. Moreover, it is more difficult to describe a CPM which is both useful in the clinical setting, and easy to use.	
120	There's so many nuances in data coding that not everything that I feel would be helpful in understanding can possibly be presented. This also makes it difficult for comparing and contrasting with previously-published papers.	

ID	Text	Theme
153	There are quite a number of smaller decisions (eg what variables to consider, choice of models, etc) that are difficult to concisely describe.	
179	The methods to creating models are variable and as a starting researcher, it may be confusing as to what methods to use. Assessment of internal and external validity are also widely variable. My advice to any researcher wishing to create a useful prognostic model would be to consult a statistician.	
182	In my field, there seems to be a tremendous amount of low-quality prediction models, particularly for prognosis. Many of these focus only on model AUC or C-statistic, which is not the most appropriate measure for prognostic models.	
204	Many do not know how to validate models well or test assumptions. ... A further huge factor is many bioengineers are not really understanding the clinical aspects well, and thus do not fully understand what they are predicting or what would make a prediction good or bad	
216	Prediction models require robust assessment so need to be thorough, therefore it is not always so straightforward to do.	
229	The AI studies I have reviewed are more difficult because there is less conceptual rigour to the ideas and they often seem remote from practical clinical interpretation.	
8	A lot of prediction model authors use very unsophisticated methods. Sometimes it is challenging to get across why more advanced methods may be needed to produce and test a reliable model.	
3	The main problem is if authors put a lot of effort into studying something that I as a reviewer perceive as less relevant.	Difficulties in providing feedback to authors among reviewers
12	it is not that difficult to write useful results accompanied by good predictive performance. However, providing interpretable models, explaining results and the interplay of risk factors can be difficult. This also affects how trustworthy results are.	
62	As a reviewer it is easy to provide a critique pointing out the strengths and weaknesses of the prediction model.	
79	i review papers from the perspective of one who writes these and that makes it fairly easy to evaluate and provide feedback.	

ID	Text	Theme
85	I do think it is similarly easy/difficult to author trustworthy/useful results, as the assessment of how trustworthy or useful a model is, is often subjective. And in my subjective view, there are many not useful models published	
85	If a model is useful is depending on the subjective estimation of a researcher. Therefore I find it difficult to give feedback that a model might not be useful. it is easier to give feedback, that it might not be developed in a trustworthy manner	
95	Knowing the pitfalls as an author helps you judge other peoples' manuscripts	
107	The basic principles that need to be checked are not rocket science. As long as you know what to look for and what key errors or misconceptions exist, then it is not hard to nudge authors in a better direction. The key is to acknowledge limitations that are not addressible in the present study.	
117	Based on a review of the study methods, it is usually apparent if the study results would not be understandable or trustworthy. Therefore, a reviewer can readily provide feedback to ensure that results are understandable and trustworthy. Whether results are useful is somewhat more subjective, so it can be more difficult for a reviewer to provide feedback to ensure useful study results.	
158	It is easy to provide feedback, but it does not entail it is easy 1) to be sure they are completely relevant given the clinical field; 2) to have the feedback accepted by the Authors.	
163	I will sometimes review such papers written by authors without even biostatistics/clinical epidemiology knowledge, for example, computer scientists. Some such authors may not even have a basic idea of e.g. challenges to external validity or causal issues	
196	Answering all of the above, giving feedback is much simpler than developing the tools. The methodology for assessing these models is usually standardized so there are only so many ways you can evaluate the models. Useful results and trustworthiness go hand in hand, if there is a solid result validation process and the results seem to improve or have any advantage over the state of the art I¬¥d say a paper is good. Otherwise, it is simple to objectively point out the faults as a reviewer.	

ID	Text	Theme
233	I have a lot of experience. In particular, as I develop tools, I am good at advising others in reviews.	
260	It is easier to provide feedback as a reviewer than to evaluate my study as an author because I can evaluate others' research more objectively.	
74	There is a lot of variation in methods, some of which matter and some likely don't. The methods are sometimes opaque. They're also changing a lot, so it is sometimes hard to understand a model from a distant field of machine learning. I'd also say that editors often don't care. I recently gave a quite negative review about a model I thought was unclear and had many hints of overfitting and the editor basically ignored me and gave the paper a straightforward R&R.	

APPENDIX F

Interview guide

This study uses the following definitions for **understandable**, **useful,** and **trustworthy study results.**

- **U*nderstandable* study results** refer to the degree to which a user is able to interpret, exemplify, classify, summarize, infer, compare, and explain the study results of prediction model research papers.
- **U*seful* study results** refer to the degree to which a user is satisfied with their perceived achievement of pragmatic goals, including the goals after reading the study results of prediction model research papers.
- **T*rustworthy* study results** refer to the degree to which a user has confidence that the study results of prediction model research papers provide information as they should.

Time period 1: U*nderstandable* study results

1. For you, what does it mean to have understandable study results in clinical prediction model research papers?
2. How can data presentations (such as text or visual formats) in the results section of research papers help make prediction models more understandable?
3. Are there ways these presentations can make study results less understandable?
4. How do you think about using data visualization to make study results more understandable?

Time period 2: Useful study results

5. For you, what does it mean to have useful study results in clinical prediction model research papers?
6. How can data presentations (such as text or visual formats) in the results section of research papers help make prediction models more useful?
7. Are there ways these presentations can make study results less useful?
8. How do you think about using data visualization to make study results more useful?

Time period 3: T*rustworthy* study results

9. For you, what does it mean to have trustworthy study results in clinical prediction model research papers?
10. How can data presentations (such as text or visual formats) in the results section of research papers help make prediction models more trustworthy?
11. Are there ways these presentations can make study results less trustworthy?
12. How do you think about using data visualization to make study results more trustworthy?

APPENDIX G

The complete list of paragraph chunk included in the qualitative analysis to answer RQ of Chapter 5, which excluded the paragraph chunk to identify visualization tasks in Chapter 6

ID	Text	# Need	Theme - Sub theme
19	Now, I think that's, you know, there's a certain standard that needs to be there so that people understand what it means. They understand the idea that they are not coming up with a perfect model. They must look to be able to always interpret or concentrate on their validation results if they've done that, and not try and present equally development and validation, because validation is where you're looking at the, closer to the real world performance of a model.	1-1-1	Understandable - Knowledge
19	And there are some other techniques, but there's also some... I'd have to go and look them up. I might have to look them up and send you them later. There's some good guidance like the trial and error, and it is like the tripod statement. There's some good guidance that helps people get some basic information that they need.		
2	So we do have layers of problems. And the main one is that there is no standardized way to show this. And this is what I've indicated in the response. We need some kind of structured guidelines that will, first of all, tell the author what is expected of him. And secondly, the reviewer or the reader can go against these guidelines and see whether all the ticks were made. This was done for other study types in the past. There is the whole equator website, equator platform that provides really this.		
3	So, it is basically like the standard reporting of the results. My experience is that when just assume that somebody is going to do some systematic review of these models, and they should have those standardized information, your results, in your at least you report in your studies. And then, so the understanding is basically related to the standardized or uniformity of the information that is already there. You should report that, and in a way that is already acceptable in the community, of the research community.		
3	Yeah, I already defined that the understandable is like the standardization of those results.		
1	How I judge I guess I'm thinking because I mainly review these papers. I don't often write them. So when I'm reviewing a paper like this, I'm looking to see that they've reported things		

ID	Text	# Need	Theme - Sub theme
	according to sort of like conventional guidance. And in some papers, you can see immediately that it is very poorly written. And they've modeled up a lot of things.		
12	And I think, in a sense, a lot of what I'm saying would relate to that, that they specify that they've, you know, there's a specific measure of discrimination and a specific measure of, I've forgotten the word, we plot it, I've lost the word, calibration. So those sort of technical terms are easily defined, standardised and should be part of reporting.		
3	And my understanding is like the, if you are following the tripod one, then you actually do not really miss those things, like not reporting the number of like how many sample sizes, how you arrive, how you arrive at those sample size, and how you actually did the modeling, what kind of like methods you use, like the likelihood method, or the other methods, if you use some statistical software, like packages, or some version of that package. So, these kind of things actually make things better, like what actually the paper is doing.		
12	OK, well, I think, for example, if there was a quality criterion for reporting these, for example, that required that there were certain measures that were consistently used, that they were explicit that they had used them and that there was a sort of agreed glossary and definition, then I think that would help. And simple examples would be, and this would relate to the, I think there are now, aren't there, you'll know better than me that there are now sort of checklists for reporting these studies.		
2	They provide checklists or guidelines with numerated items where you kind of have to follow the flow. Otherwise, you cannot say that you abide, that you are, it is basically a benchmark of what you have to do in a study. Without it, there are inherent holes or flaws in papers, and it is difficult to assess. I mean, you read it, it sounds nice, but you don't know what it means.		
13	And usually they talk different languages. And it makes it sometimes a bit difficult to formulate or describe or present the results because statistician looks on different points than a doctor looks on. And therefore, it, so you need both competences to have a good prediction model. Even in planning development of such a model.	1-2-1	Understandable - Comprehension
6	I think that there's kind of, I think about understandability, both from the clinician side where or the end user of prediction tool, and also from, you know, more another technical expert or statistician understandability.		

ID	Text	# Need	Theme - Sub theme
19	Okay. They need to have a statistician involved with interpretation early on. Just presenting a ROC curve is not enough. And so there are... I'm actually writing a paper for cardiologists about this kind of thing at the moment.		
19	If the person (statistician) is part of the team from the very beginning, it can be better because they get, the physicians get a better understanding of what kind of data they need, and how it goes together to form a model.		
13	And therefore we have sometimes problems that, so if the statistician writes the results into a paper, then he cares for terms or expressions or presentations so that other statisticians can see what he has done. And this is sometimes a little bit hard to understand for the doctor. And the doctor would maybe use other words to explain the same thing.		
13	Okay, so clinical prediction models is a, there are two disciplines involved. On the one hand, it is the clinician, he has a certain prediction problem or would like to have a prediction of something. On the other hand, if you want to apply not the very simple statistics, you need some knowledge in multivariate statistics. So you need on the other hand, an expert in statistics or in modeling. And this is, it is very seldom that this is within one person. So it must be a cooperation between a doctor and a statistician.		
13	And usually it is some journals have it at the end say which author did contribute in what aspect of the paper. And from that you can usually read that there is a statistician among the authors. And this is an important point.		
9	So clinicians sometimes will not use these prediction models even though they may be backed by data or what have you because they don't understand, like I said, we just talked about, they don't understand the model itself. They don't know how it works. So it is like a black box.		
6	So I think it is important for authors to consider the end users. So if they don't understand it, then they're not going to use it. Although, as I kind of, I think mentioned earlier, I don't think that they need to, you know, if the inputs track what they think is important, I don't know, it is going to vary from person to person. I don't know how many, if everybody needs to, is going to care about how the sauce is made, so to speak.		
2	what a mathematician sees or what a modeler sees is brutally obvious to him or her, and completely not understandable to an average reader who wants to go and see whether the study has some merit. Well, KISS, keep it simple. If you can have a predictive model that is		

ID	Text	# Need	Theme - Sub theme
	functioning, then try to explain it in the simplest possible terms. Otherwise, you may miss in the end.		
10	And I suspect that most people that use or read clinical prediction modeling are not statisticians and they want to understand what's the gist of it, what's the point, and understand is the key word there.		
12	This ought to be, let's take a measure of discrimination that, you know, that, that, a C, a Harrell's C statistic, for example, that ought to be completely and easily understood by any healthcare professional. So I would say that needs to be very, very clearly explained with an example in the text.		
17	And then the results should be explained in both these terminologies. What does it mean to the clinical fraternity? And why should we take this with a grain of salt in that what are the limitations that's coming in from the mathematical model that's embedded into my results?		
12	But most, a lot of a huge number of a huge number of clinical prediction papers are, say they are aimed at clinicians, at healthcare policy makers, and I don't think that you've picked up the point that I often don't think that they're taking account of that.		
7	So, I would say that, yeah, to my first, the clinician, or the, the, the clinicians or the healthcare professional will be using it, using the model. So that's one, one, one part of the stakeholders and the second one was the patients, or the. Yeah, the persons for whom the prediction will return a result. So I think both of them. I would say so.		
18	But you know, from the risk prediction perspective, you are better off having the prediction as a continuum. So you have those sort of different ways of reporting, but neither is complete. I mean, you could report discrimination and collaboration and then somehow find the perfect cutoff point and then report sensitivity and specificity and all those things that really matter for clinicians. But that's not that important from the risk prediction modeling point of view.		
12	And I spent my life, you know, puzzling away at these issues and of course what the latest round coming from the machine learners is, which I think is interesting, is they say don't worry about that, that's just noise and, you know, if we see patterns we're fine. But still for me that notion that you know what you're dealing with, and it was interesting		
19	PROBAST, yeah, there's some other ones as well, but there's also, a new one they're working on, for machine learning, as well. They're helpful, but, sometimes, they're just people ticking the boxes. I've seen people tick the box, saying the title, obeys things, but it doesn't. And, yeah, one		

ID	Text	# Need	Theme - Sub theme
	of the things that, that if people haven't done this kind of work before, it is actually useful, as a starting point, those things, before you do write the paper, before you do the research, even. To look at some of those, minimum standards of how to, present things.		
9	But it is just, so like just from interviewing clinicians, what they find is a sticking point is that a lot of times these models seem like a black box and a number of them have used the term black box. Like, what are the inputs going into this model among other things?	1-2-2	
9	But sometimes some models seem mysterious, right? As to what the inputs are and what the model does with the inputs, you know.		
12	I think what would help understandability is a clear statement of the point in the progress of their study, where, you know, that has gone into, you know, the black box and what the method is that has been used. And some easily understood brief explanation of what that might entail, you know, whether it is, you know, this is clearly a method where the computer has been taught to recognise certain patterns, or whether this is a machine generated model on the basis of the information that's there, just something that's a bit more explicit, rather than what a lot of the papers just assume that, well, we've taken these variables, we fed them in, and out they come at the other end. So a bit more explicitness, I think.		
4	And this is more and more going to be relevant, because some of the prognostic models are going to be using deep learning methodology, which is getting exorbitantly complicated. So, the idea that reviewers of clinical research models will understand the intricacies of the model, I think is getting less and less likely. But they will still be able to interpret the results. I don't know if this fits into your framework, but that's just what I think.		
6	Now, understandability for the end user I think it is kind of tricky, because the question of whether, you know, I think of, is it possible for black box algorithms to be understandable or not. And depending on your view, the answer to that might change because is it understandable, if the person knows what goes into if somebody is familiar with the inputs, and they know sort of why they matter or it makes sense to them why they matter.		
1	So that's at the one extreme. At the other extreme, I think that the papers tend to be well done. So they've got like a nice baseline table. They've used the right type of regression approach. They've used multiple imputation if that's appropriate. They've looked at nonlinear effects of covariates internally validated. So they followed all that sort of guidance. And then it looks like they've got a nice prognostic model.	1-3-1	Understandable - Analysis

ID	Text	# Need	Theme - Sub theme
8	But I think that there needs to be a clear explanation of data cleaning procedures, normality assumptions, things like that, kind of data checks that make it very clear like, this is the type of data we're working with. Here are the things we're assuming about it.		
17	So in terms of authoring this, I think it is very important to show all the groundwork that we do in terms of data cleaning, data pre-processing and checking the data for all the assumptions and then fixing, like there are some, you can do a test for normality, for example, and then you can apply some transformations to make sure that your data is correct. It is difficult to author this because like I said, word limit and figure limit make this very unappealing and most of this just gets summarized in one single sentence or two sentences at best.		
17	I think this goes hand in hand with understandability in what we discussed before. You know how we said data visualization at the early on stages of data discovery, if there is a mandatory figure instead of having just a summary table, then we can show that this was the raw data, how it was found, it was not normal. I applied this transformation and now it is normal.		
17	So instead of, like, you know how clinical papers we normally have introduction, methods, results, discussion. If clinical machine learning papers had introduction, methods, results, and the results section had a couple of mandatory data discovery figures that were required and the discussion section went into understandability, explainability and reliability or some of these main concepts that were required by the journal to touch base on, I think our papers would have a lot more quality overall.		
7	I think I'm not sure it is common practice for prediction models. It is common practice for trials, or things like that. But for prediction models, not that often say, but you're suggesting to use that. I think so yeah, yeah. I see some sometimes, and I think it is very helpful to know what what data is used, how many, what's the sample size, how many outcome events, etc.		
17	Sometimes, this may not be the case that is relevant in terms of how it really helped the model. For example, if it is a case of linear regression, I find most clinical papers do not go into issues with multicollinearity, which is a very important aspect about linear regression. You want to have all your features being independent.		
17	This is one thing that I don't see used often. Normally, people put a table that says this is participant, how many participants were male, female, what was the age. If it was a stroke paper, then they normally say how many milliliters of stroke there was. And I have also done the same thing in my PhD projects, is you have a table that summarizes patient characteristics or		

ID	Text	# Need	Theme - Sub theme
	what the data has. But beyond that, I feel like as a field, we don't go beyond to try to come up with data visualizations that show more properties of the data that helped us build those predictive models.		
6	You know, if you perturb the data a bit. You could get a completely, completely different simple model. You know, if you say I'm only want to have to, you know, some, this is a contrived example but you know if you say that you only want to have like a maximum of three predictors in your final model, and it has to be some sort of score, you know, like integer scores coefficients. If you change some of the input data, because some predictors, you know, may do as so like I guess I'm thinking of maybe if I give a concrete example. You know, using weight versus BMI, or that's, you know, not exactly but it is possible that you could have two models that predict equally well that have three only a small number of covariates that are completely felt like the over, they don't overlap.	1-3-2	
7	And then maybe the, well I would like to see more visualization, maybe the design, like the design of the study, the analysis, what it is exactly what kind of data, how the model validate it, like, is that external validation or internal validation. If there was internal validation, like what kind of technique. So I think all that, and I think the flowcharts do a good job. Yes, yes, how it was conducted, like, what is that purely developing a new model is that development, plus validation, what kind of validation what kind of data, etc etc.		
13	For prediction models, I think you should know how the prediction is reached. Which factors are included, what is the weighting of each factor and so on.		
12	If I'm at the end of it, I'm satisfied, I'm satisfied that they clearly, they've got things like the outcomes, and the way that they've selected variables clear that background, those background processes. So I'm understanding, you know, whether they, they selected it, the variables on clinical grounds or on statistical grounds.		
14	It is just, to me, it seems like just good research, you know, components of good research that, you know, your goals are very well stated, your, your variables are, are very well defined, that your methods are well defined. Whatever, you know, how did you, how did you select these variables? What were your, your, you know, what was your variable selection procedures, your validation procedures, all that kind of stuff. So all that stuff that needs to be very well defined.		
2	Just like I said, I've seen a lot of papers and plenty of them finally do end up with some kind of data or results that seem very reasonable. However, what we don't know is how many attempts		

ID	Text	# Need	Theme - Sub theme
	were failed and what these guys went through to get to the model that is functional. And then occasionally you see horrific things saying model works in elderly women who don't smoke and who have three grandkids, which is basically it means that they've managed to pick up part of the model that for some reason ended up predicting it well. But in reality, there is no context in which that makes sense at all.		
12	And I'm clear about where that sits in the panoply of development, validation, and external validation, and so on. And what generalizability this might have. And you could say that that can all be in a sort of hermetically sealed thing about the explicitness and the clarity of the research bit, the statistical scientific bit, we need to understand these variables, how they've been selected.		
9	Yeah, like I said, it is been a while but basically the only thing I can think of right now is what are the inputs? Just what are the inputs, for example? You know, how are they weighted? I don't know, I'm not a statistics person, right? Just what are the inputs and what happens to the inputs in the model and then you get an output, right? So this stuff right here could be explained better too. Yeah, yeah.		
1	Then they will summarize the performance in terms of the C statistic. But then mostly they leave it at that. And they end up concluding something like this is a prognostic model that is useful, C statistic 0.8, 0.9, maybe 0.75. And their overarching comment is that this is going to work, probably recommends that it be used.	1-3-3	
8	Yeah, like I said. I think that having examples of the predicted model, maybe figure one is the theorized predicted model, and figure two is the actual results. So that we're being able to compare the difference between what we predicted and what we actually saw.		
1	And then you use natural frequencies to show what would happen to typical 1,000 patients if this prognostic model was used on them. How many of the events that you would detect, how many you would miss, how many you would miscategorize. I think that's the best way of doing it with numbers.		
7	And second, the performance, how, what does it mean to have a good calibration, or like the calibration curves show some show some patterns, then one needs to be able to understand that. And that's what it means in terms of calibration or other metrics. So that's, that's what I would say.		

ID	Text	# Need	Theme - Sub theme
14	And then, oh, the graph where you're graphing the predictions against the outcomes that shows you the accuracy of the predictions. I think that that can be helpful.		
20	Um, I don't know. I think it is just the things that they omit that make it less understandable. That is to say, if you don't include a calibration plot, people may not understand that the model won't be calibrated for their own data. So I could, I could show just the discrimination in a, in another dataset. You know, I could show that I validated it in an external dataset and that the discrimination is similar to what I saw in my original dataset. The discrimination might be the same, but the, but the calibration might be really different. And so, you know, unless I know what that is, I can't use the formula that they derived to come up with the probability of somebody having that outcome in my data. Okay.		
11	Well, certainly, so all the graphs that would show comparison of expected and observed outcomes, or if it is a, if it is a score, you know, the area under the curve helps sometime understanding how, you know, whether it is more like a sensitive or specific or, you know, things like that.		
14	But then when you look at the performance metrics, then maybe a model that doesn't make biological sense, or you don't understand why an individual variable would be split in this particular way, or whether the spline came out this way, but it still looks like the model is performing well. So we think of the C statistic in terms of discrimination, or we think of Breyer scores, or we think of, what am I, I'm thinking of that graph. I'm thinking of the accuracy of the model, any indices that we use from that. So we might have models that statistically don't look understandable, but then they look like working okay. So that's, you never want that, because then you don't know. It makes you a little suspicious.	1-3-4	
3	But yeah, there is no really specific… Yeah, it is interesting that there is no specific, like, style. But it depends on the problem. Like if you're working with some problem and then, and also your outcome variable, like what kind of your outcome variable and how you can use the… Like in my experience, like I was used to work more with the continuous outcomes and that there is a standard regression analysis or maybe the GLM like Poisson or the other regression models are useful. But when I start working with the binary outcomes, which is a totally different game, so you have to be a very sure like the how you define the relationship of your covariates with the outcome.		

ID	Text	# Need	Theme - Sub theme
14	So when, well, how can authors overcome that. So you have to be able to, to put, to put what you found into plain English, you have to be able to explain that other things going on behind the model of variable that seems like it shouldn't behave the way it is makes it behave that way in terms of, you know, how much variability there was in it relative to other other variables in the model or whatever it is, you know, it is even hard for the statistician.		
14	I think about understandability statistically, we can create models, but if they don't make sense, they might not make biological sense. They might not coincide with a conceptual model that we have, or you're sitting there and you might think, I do not understand how this variable would affect the outcome in this way, or maybe what I'm seeing in the model in terms of the weights and whatnot are counter to what, how I thought it would turn out, those surprises come up. And then depending on what type of modeling you have, you're doing, if it is recursive modeling or whatever, you might come up with things that just don't, they don't make sense.		
9	But yeah, I guess transparency about the model itself and how the, whatever the output is, how that's calculated would be useful to them.	1-4-1	Understandable - Application
5	Actually, figures to make it more communicable to readers. For example, for studies of COVID-19, how to calculate the model and apply the results. Those figures are necessary. Is this okay?		
20	It is just a one kind of a easy way for me to communicate, hey, I have a really good model that has AUC of 0.87, therefore you should use it. But nobody uses a model across the entire range of the AUC curve, right? They're going to use it at one particular point. They're going to have a cutoff and they're going to treat people above it and not treat people below it.	1-4-2	
20	And some discussion of how the model would be used in clinical care. You know, what part of the calibration curve are you actually looking at? Do you want to treat people who are very high risk? Are you trying to exclude people who are low risk? You know, what exactly is it that you're planning to use this model for? And then just saying, we have a model to identify risk factor. You know, we've identified risk factors for a particular disease.		
8	And then on the flip side, when reading clinical research and digesting it, I think that a good writer will make clear what the actual clinical implications of the results are. Sometimes I think in clinical prediction models, there's not a very good communication of how these results apply to a real-world situation or what they might tell us about something useful that we can really take hold of and apply in a clinical setting.		

ID	Text	# Need	Theme - Sub theme
12	And so many papers aren't absolutely clear about where it is going to be, where it might be used, what it might be used for, what's the intention, and this will go back and drive, you know, the selection of prognostic factors for a model might be entirely on the basis of, might be entirely on the basis of the fact that they're easily measured in practice.		
18	And not only in terms of like statistical methods or numeric results, but also in terms of like, as you show the results, what do they mean? And what you can do with that and what it may mean for your patients or for the people with whom you are expected to use the risk score in the end.		
19	And then even if they've tried to do some statistics with that, they haven't stopped and said, how are we going to use this model? And is it really better in the use case? Which is usually at one, at either very high sensitivity or very high specificity or both. They might have two or three thresholds, which is where you need to compare it, if it is going to be clinically useful.		
5	So measure of the outcome. So the easiest one will be, for example, mortality rate. If the model could predict a mortality rate, it would be easy to understand. Understand? Putting some clinical variables and a model estimates a mortality rate. And it would be easy to be applicable to clinical practice. But the odds ratio is, will be difficult to apply to clinical practice. Why is that? Why? Because it is not risk ratio or absolute risk. I think it is a common understanding. Odds ratio is difficult to apply to clinical practice. But it is very useful for meta-analysis or something.		
19	The, with clinical models, depending on whether they are sort of a screening model, if they're screening, or whether they are sort of diagnostic and trying to get a high, I don't know, positive predictive value or something. Sometimes there's sort of graphs that show the proportions of paper of patients who might be ruled out or ruled in, and certain thresholds, might be a probability threshold. From a clinical perspective, that's of interest, although it is influenced by the prevalence, which you've got to be careful with.		
7	To me to have understandable results mean that at least we understand the domain where the clinical prediction model can apply and can be potentially used. It means that one can understand by by domain of applicability. I mean, one should be able to understand what is the target population or the clinical fields where it is meant to be applied.		
11	Yeah, I think so, certainly, certainly the, what the, you know, the variables, right? So do I understand the predictors? Were they clearly defined? Do I feel that I can reproduce that when in my patients and then in the rule as a whole, you know, does it provide me with an estimate of		

ID	Text	# Need	Theme - Sub theme
	disease, probability of disease or something that they can, that they can actually use as well, right?		
14	From an end user point of view, so I do, my audience is orthopedic surgeons, pediatric orthopedic surgeons. So when I write up what I have done, or when I'm reading something, it needs to be understandable in terms of clinically, can they see how this, what this model might mean for their practice, how this model might be used, how they might be able to, how they can explain this model to their patients.		
10	So that's usable, but not understandable for me. Understandable means what you used, what you see, you understand and you integrate rather than you use easily. So you integrate the information and you apply it in clinical practice without necessarily always seeing the figure, because you can know that, oh, I remember that whatever the severity, the more the biomarkers are high, the more red, the more bad it is for patients. So it is like the Framingham risk scale, which of course is usable in a calculator, but everyone understands because they saw it in a figure. And that would be about understandable.		
17	And the same thing applies to adaptability as well. So this is coming from, if I read a paper that's not in my field, so I have a problem in neuroradiology but I read a paper in cardiology, then can I adapt that solution to a neuroradiology problem? So is it generic enough that I can apply it to within my field?		
14	Important because I think that they can they can grasp this idea that if you say, and this is what they preferred us to say to them out of 100 people like you. If they did not have treatment 5050 out of 100 of them would end up having surgery. So that was kind of a preferred interpretation that they wanted. I think that that's pretty personally true that people like that, that kind of verbiage of 100 people like you. This is what would happen, and people can tend to understand that.		

ID	Text	# Need	Theme - Sub theme
14	I've actually tried to tackle this problem, not just with the clinicians but also with patients. And so, my patients, the group that I've been mostly working with our adolescent girls who have scoliosis, and so I believe that if you can give them an estimate of what their risk of of a certain outcome. So we say, somebody with your characteristics has a 98% chance of having surgery, if you don't do this, then how do I explain that to them. And so there are various graphics that you can use sometimes people will use the, the, you know, hundred people and you know, all of them are read except for two of them and that would indicate that there were two people out of 100, you know, or line graphs or paragraphs and so I actually did some surveys and focus groups with kids and their parents saying, which of these types of graphics make the most sense to you.		
20	I think that they lack even the most basic understanding of risk and of discrimination. So that if you say to them that the area under the curve is 0.85, they know that's good because somebody told them greater than 0.8 is really good. 0.8 to 0.7 is kind of in between and below 0.7 isn't very good. But they don't know what that means. They've just translated a number into a word. They have no concept of how would it, if I applied this to my patients, how would this change my practice? How would this, you know, would I be better at differentiating between patients who do and don't have disease?		
8	Well, I think that I'm a clinical psychology graduate student. So my opinion is that the best research is informed by clinical practice. And the best clinical practice is informed by research. So I think that it is really important that when we're doing clinical prediction research, that we have a bidirectional understanding of what's going on in the science practitioner sense.	1-5-1	Understandable - Synthesis
7	I think ideally yes, I think that's, I think that's one of the current challenges of research in general, is that gap between those who provide information like the researchers, and the targets or the people may be concerned by the use of that. So I think, I think yeah ideally, I think, I hope that in the future we will see such papers, which will include patients or stakeholders or, yeah, I think that would provide a much, much more holistic picture		
7	In addition to what I said, I don't think so. Maybe my last comment would be to underscore the fact that the end product of what we want to achieve at the end of the research should always be kept in mind. Like, for example, providing clinical prediction models that is useful and that's improved the patient's outcome.		

ID	Text	# Need	Theme - Sub theme
6	So, in conclusion, understandability is important both for domain experts and end users, but there are trade-offs between simplicity and accuracy. Ensuring that a model is understandable may require considering how it performs across different subsets of data and the potential implications of simplifying the model too much. Balancing these factors can help create a more robust and useful prediction model that is easier for both practitioners and users to comprehend and apply in real-world situations.		
8	And so I think, like I said earlier, the authors should draw a clear line, tell a clear narrative of what this result, what line of research this result is contributing to, and what the overall goal of that type of research might be.		
8	But let's say that you're using an experimental paradigm to demonstrate something important. Well, I think a good clinical prediction paper will take the next step in the discussion section to really paint a picture for the reader of what part of the next part of research this will inform and what the overall goal is. So is the overall goal to create more tailored treatments that are culturally, contextually specific to an underserved group? Well, paint the picture of how we're going to get there and kind of make that statement, even if it is just in the last portion.		
13	I'm not really got the point whether the author should show what their contribution to the paper is.		
14	So I think the first thing to do is to not you know, if you're doing, if it is for research purposes only, and you're trying to explain what's going on, that's fine. But if you have, if your purpose is to have a positive impact on clinics, then this has to be something that's, you know, user-friendly, clinic-friendly, right? So I think that's the first step.		
14	So I think that in addition to the statistical derivation kind of invalidation work that we have to put into the paper, we also have to provide them with information about what the implications of use could be. I think, you know, those things all, all fall under understandability. Do they understand how this could be used and what the implications could be for their practice.		
8	So I think that on one hand, the research that's being conducted needs to be, I think, a good clinical prediction model is informed by practice-oriented perspectives. So not just looking at, let's say, for example, the efficacy if we're talking about treatment. Not just talking about the efficacy of a treatment, but also examining the effectiveness or taking into account real-world problems that might be affecting the clinical picture.		

ID	Text	# Need	Theme - Sub theme
17	So a lot of these journals are already mandating code sharing and reproducibility and generalizability of machine learning models.	1-6-1	Understandable - Evaluation
3	And about, of course, the setting of the data, it should be transparent as much as you can, and then this could be doable. Like if somebody is replicating the results, are you doing some similar study, so they should be able to actually replicate. So, the main agenda is to be like these models are generalizable as much as it can, but reporting is standardized.		
2	One suggestion that could work is for you to suggest that authors need to provide both raw data set and the analysis flow alongside the manuscript. It is claimed that other people can then access and use it. In my experience, you need to have somebody who is emotionally very against you, who hates your guts to go through and follow your steps. But still, the fact alone that you have to attach the data and attach the analytic protocol means that you are probably doing a better job than if you just did something yourself and then said, oh, that's how it is. So maybe attaching that would be some kind of almost an indirect element of quality, but it kind of pushes you to be more careful, to be more structured.		
6	So, understandable to me from I guess the easier side to me is the technical side of what understandable means. And to me, I think a basic requirement is that another person, another statistician can both understand or like follow your justifications for your methods, and like it should be clear why you chose what you what you chose both, you know, method wise and performance metric wise. So those decisions should be understandable to a domain expert and sufficiently detailed ideally with code available that somebody could reproduce the results, if they had access to the original data, which is not always possible but you know I think that's ideal situation is that both somebody would be able to replicate the code on the actual data that was used.		
7	And then the second is for sake of reproducibility. Of course the coefficients of the model, how to calculate that should be made transparent.		
3	Like, for example, I mostly work with logistic regression, and some people like report like area under the curve, some use maybe the Breyer score, some use maybe the other performance measures. But in medical literature, that is commonly accepted that you should have at least area under the curve as a performance measure. But if you say that you have not reported area under the curve, it is a pain in the neck. So, if you are publishing in a medical research, it means that your results are not really comparable to other studies and not understandable.		

ID	Text	# Need	Theme - Sub theme
6	It is a little tricky in my view because, or I don't, I don't think simple models are good. Sort of, but they're also. Sometimes, there's all kinds of trade offs both potentially accuracy versus understandability but even if something's understandable. Sometimes I think that the process, understandability can trade off maybe with clinician understandability, in the sense that. What I'm thinking of is that sometimes these methods that create very simplified models. You know, if you perturb the data a bit. You could get a completely, completely different simple model.	1-6-2	
10	So I think, of course, it is a trade-off, because between understandable and robustness, of course, an equation, a regression equation will always be better. When you do categories, when you use a table, when you cut off points, you always lose a bit of finesse. But this at least is understandable, and you integrate the information in such a way that you do not need to calculate it all the time. So I think that's where understandable becomes very important.		
20	Most of the papers aren't going to include a cutoff. But if they do include a cutoff, they're not going to tell you how that's going to influence your care. You know, what are the outcomes of your patients? How are they going to be changed by using this model or not using that model? Those pieces of information are never in those papers.	1-6-3	
20	They're not Bayesians. They don't understand how is it going to change my pretest probability to a post-test probability. And so it is really impossible for them to understand these papers. And I'm not sure that the papers could be written in a way that the physicians would understand other than to compare use of the model to clinical practice.		
20	And so when you say understandable, it really kind of depends on what level of understanding they need to have. So for example, many of the prediction models will use an area under the ROC curve as their way of comparing their model to another model or stating how good their model is. But that doesn't really tell you anything about how it would be used clinically.		
19	And so they're not looking at what is really plausible. So that means the interpretation, you know, I've seen interpretation that says one model is better than the other, when they've just compared the estimates, and nothing more of the AUC.		
19	So, you know, for example, I do a lot of work with, I do a lot of work with, in the emergency department. And with chest pain, we now are able to use different algorithms to rule out people having a heart attack, myocardial infarction, with very high sensitivity at a certain threshold. 99% sensitivity is where we work. So we can rule them out. And that is, you know, that's the clinically relevant threshold for that particular condition. Now, you could have two models with		

ID	Text	# Need	Theme - Sub theme
	overall AUCs, you know, one better than the other, but at that threshold, different, you know, it could be the other way around, which behaves better. So you talk about interpretation. There's two things. One is a lot of models, and a lot of, unfortunately a lot of clinical science is interpreted wrong from a statistical perspective, but also from a clinical perspective, what is really useful.		
7	And I wouldn't say much about the explainability of the model in itself, in the sense that to me prediction models quite differ from causal inference models. So, one should not try to interpret the example the coefficients, in turn by prediction models, because by the by by nature, like they are. They're biased for causal interpretation.	2-1-1	Useful - Critical support
7	But I think it should be should be made transparent in a way that it wouldn't suggest causal interpretation. I'll give you an example is, for example, if we do run a logistic regression to predict a to to have a prediction model of a binary outcome. I think the coefficient should be reported as coefficients like beta coefficients, and not as odds ratio. First, because the putting it in odds ratio scale will reduce the precision, like, because it is just because of the units you start to round.		
13	Therefore, I kind of hate this kind of machine learning or black box systems where you don't know how the prediction is calculated. You have to use this black box system which others do not have. I couldn't apply it myself. I have to use the system from the people there.		
13	What is the aim of this project? The aim is, in fact, you should think about how to really apply such a result. If you calculate a prediction model, it is nice to know which factor is more or less predictive. This could be an aim.		
13	You should be clear when you start such a project, what for, what is the reason, why do I do this? The system that I have developed will be used in quality assessment. We have an expected outcome and we have an observed outcome and compare this to the system that I have developed. This is something where we apply such a system in retrospect, so we do not apply it in the emergency room. After work has been done, we calculate the expectation and compare it with what has happened.		

ID	Text	# Need	Theme - Sub theme
10	And then the second thing is for me, a risk prediction model works on modifiable risk, which that you can not enact upon. Of course, if you're measuring, let's say the asthma attack risk, historically, people would say you had an asthma attack in the last year, you're going to have another asthma attack next year. That's the main risk factor. But you can't do anything on that. You can't come back on the past. You can't do anything. You just say to the patient, well, too bad, you're going to have another attack. Same for, I guess, heart attacks. If you had a heart attack or you have angina, you have a higher risk of heart attacks.		
17	And so what, like, why is it so important for someone who's trying to schedule a surgery in 20 minutes? What is the real benefit?		
2	And then, of course, you need the substantial limitation section saying that this is all based on the data that we have, and it will only work under the assumption of no major contextual change in the system, because if the system changes, then the entire model kind of goes away. So I guess that the post hoc explanation of the model is something that should be developed more.		
3	And the last thing is the is it about their clinical usefulness using the now we have a decision curve analysis, which is like is inspired by the economic analysis. But it is just to understand in terms of false positive and false negatives. What is the trade of this model is like this is going to harm the society. This is going to either is a better for the like there is some on average benefit of using this model. Are you just as is a standard to the like the same practice that they are using?	2-1-2	
3	And if you not treat anybody, if you treat anybody, all the person you treat with some thresholds. So what they did is nothing was like in this in limits. It was actually going beyond the limits. It means that the model is actually harmful. And but the author hasn't really interpreted that well to communicate that this model is actually useless. It is just rubbish.		
1	But if you embedded this in the entire 60 million population that live in the UK, it is going to be flagging up lots of people that are at risk. And then how do you deal with that? So it might be sensitive, but it is not very specific. And then if you, if you use it, you're going to be flagging up lots of people being at risk.		
1	And it is just exactly what the first person was saying, that there are these prognostic factors, but actually many people have those factors and they don't go on to commit suicide. But when you put that into an AI model or a prognostic model, people seem to fail to remember that that's still happening in that model. And then you add that on to 60 million people and maybe you		

ID	Text	# Need	Theme - Sub theme
	said that 10 million were at risk. Well you couldn't cope with 10 million people being at risk so it doesn't. It is not as good as you think it is.		
1	I think if a model is going to be useful, it should be accurate. But accuracy isn't just a single statistic. It is about sensitivity and specificity. .... But nonetheless, the model has to be sensitive and specific and the trade-offs have to be clear. And sometimes it is going to be acceptable to trade one off and in other cases, it will be acceptable to trade another one of those things off. But almost always, it is not clear what the trade-offs are. And I think people don't realize what the trade-offs are. So I think almost always people think that models are accurate, where actually in reality, I think a lot of them aren't.		
11	Whether it is a diagnostic test and then it could be a Fagan's nomogram of how such a pre-test will be influenced by the result of the diagnostic test and what will be the post-test probability of disease or some discussion of risk benefit of treatment and above which probability of disease the chances is that patient will get more benefits or harm from being treated for that disease.		
13	On the other hand, if you really want to apply such a model, that is, if a patient comes in and you apply a predictive model and you have a prediction model, and finally you find out that this case only has a 10% survival chance, so then do you think you want to apply such a system, which might lead to not treating a case because this case has no chance? This is a rather critical and ethical thing to consider.		
6	So, that's like a prediction goal, you want to know, you know, you're not telling them to change their medications but before they got COVID, they have COVID, and you're figuring out how they're going to do. So if somebody takes the model and says, Oh, I want to know what I can do myself to change my COVID risk. They shouldn't be using this, this tool, and this I think plays out in all kinds of risk prediction domains I know like the, there's some cardiovascular ones and there's, you know, a whole world, you I'm sure you know more than I do, actually, of risk prediction domains, but it is very tempting. I think sometimes to you know you have the tool or a website or an app and you say oh what if I did this, when they have some behaviors that you can change, but they're not designed to answer that question and they'll give biased answers.		
6	I think that at a point, you know, a good risk prediction tool if you really want to get it out there in the world and you're not just trying to say identify some of the biggest risk predictors, then it is really a whole.		

ID	Text	# Need	Theme - Sub theme
6	I was just saying about it both needs to be like simple to use or at least easy to use so that whoever is supposed to be using it actually goes through the effort of using it. But it also needs to make good predictions that's what I think that's what I meant that somebody might not use it, or like the two big reasons for not using it, or probably, you know, it just sucks to use subjectively for, you know, if the design is if the like user experience design is bad. So they don't use it because it is annoying to use. But also if they think that it is not making good predictions, then we'll stop using it because they'll think why bother.	2-2-1	Useful - Easy to use
18	Make a tool, something, whatever that is, you can actually do for people to actually use it. Otherwise, it is just going to be difficult for your model, for any model, to hit the practitioner's desk or anyone at the desk.		
18	Yeah. I don't know if I will use the word appealing, but I will use the word useful. Useful. Or usable. .... You can do something else like a risk chart. You can have it print or some tool that actually makes my life easier as an end user.		
19	So the other aspect of usefulness is, and I've seen things like Shiny apps or things on the web where people can put in the information and get a probability or a graph.		
19	But I do an icon array where I've got little pictures of people and however many, depending on the probability. Now, some clinicians go, oh, I can see that		
11	How, you know, how would the score be used and or yes, or how would they combine it with other information or yes, what are the potential consequences? I'm not saying that they have to know right away, but at least have some ideas of what could be the use.		
6	I think that there's kind of two big results I think of one is just sort of identifying the largest risk factors in terms of like the magnitude of their effect, something like that. ... if you know what the risk factors are that can suggest future research directions.	2-3-1	Useful - More accomplishment
14	So sometimes it could be useful. So you might create a clinical prediction model, and then that goes on and spurs additional work, or people start then paying attention to other aspects or characteristics of patients that they wouldn't have paid attention to before or something. And so there can be, maybe it generates more research. So there's this kind of short-term useful things.		
14	You also need to ... I think. You know, can somebody, can somebody read it and then it should be written in a way that it can spark new questions or, you know, get somebody thinking		

ID	Text	# Need	Theme - Sub theme
18	Well, I think it is if you do not report everything. That's just not useful at all. Yeah. If you don't have the resources at that time, the patient to do a tool, that's all right. I mean, that happens. But at least report everything and just hope for someone to take that information and make a tool.		
19	You have to have the first study that raises the interest, but you have to have the other studies to show that it is going to be robust if it is translated anywhere. Does that make sense?		
19	What it means to me is that it would either stimulate me looking at the same kind of data in my own data, if I'm reading it somewhere else, but I would only stimulate that if I see, aha, that will tell me something new about the patients, and maybe I can improve the current pathways of care. So that would be something that I would see as useful, not just a number.		
13	They have a model and see this and this. But they don't think about the next step. How will it be applied? Is this and this really important or is it not important? This is a different question. You don't want to apply it to the treatment.	2-3-2	
11	Or again, papers who do their prediction, but they don't say which course of action will ensue or what should we do with their prediction. I find less useful.		
9	All I can say is that, I would just put that as a, if authors want people to use their model, they need to understand the conditions under which this will be adopted.		
10	Yes, I think so. Because as I mentioned, when you have a nice figure that you understand, well, you can integrate it in practice. So one example of that, Ivan, is that, you know, the Framingham risk scale, where you have high blood pressure, cholesterol, red patients, green patients, yellow patients.		
20	Yeah, I think that they should give some idea of how you could develop these tools. And ideally, if they develop their own tools and then test them, because that's the part that's completely missing.		
7	If it is purely statistical performance in terms of prediction, I'm not sure that usefulness is completely achieved because it doesn't say just one small piece of the total puzzle. It doesn't say whether the use of that prediction model would actually lead to improving decision making and thereby the patient's outcome.	2-4-1	Useful - Quality increase
2	It depends on the perspective. If I am the end user, I would want to see the paper that assures me that this is valid, because as a physician, I'd want to know whether something will work with my patient or not. So I would like to see no major flags in a paper.		

ID	Text	# Need	Theme - Sub theme
6	I think so. I think that in some ways it is proving that your results that your model generalizes, but it also is related to implementation there, because you. And so I think that having like an RCT would be like the last stage, almost like you first need to show kind of your first you know your standard results that like your model does a good prediction, and then maybe there's some implementation science or implementation stuff that shows like, hey, we actually have a something that are like usability design like oh this is actually usable. And then kind of the ultimate step or ultimate proof would be, if you actually did an RCT where the intervention is, we use this risk prediction tool or not.		
7	I think, again, maybe it comes back to the first part of our conversation that the stakeholders for me are the clinicians and the patients. So, when we talk about usefulness, those two parties should, there should be an impact there.		
14	Okay, well, so a useful study result would be one that ends up, and sometimes it would take years, but actually improving outcomes of patients, right? Because that's the goal. No matter what we do, that's what we should always be thinking and remembering that we're trying to do.		
14	But I think useful really means that can only be decided after you've done an impact analysis. So it is only useful if it is changed somebody's behavior. And so it is changed somebody's behavior and that ultimately patient outcomes have improved. And rarely ever, never, hardly ever do we get to that.		
20	You know, if you there are like thousands and thousands of models, there are limited numbers of models that have been tested or validated in practice, and almost no models that have been shown to actually change care and patient outcomes. That's what we'd like to show. We'd like to show, hey, if you use this model, then your patient outcomes will be better.		
10	I know the third part of your talk will be on robustness, statistical robustness, but I think that usefulness is also part, it is intrinsically part of how robust the model is. Is it something that was seen in one small trial of N equal 50, or is it seen across multiple trials? What's the quality of evidence? So I guess that will, it is intimately linked.	2-5-1	Useful - Useful for task
13	For me, it is also important to have a, let's say, clear description of how the model works. And this is different if you apply some technologies of machine learning. You put some data in and some connections were done and at the end it comes a probability prediction.		
19	But what we need to make, in order to make a change, we have to be convinced that whatever the model is, that it can transfer to our hospitals, that however it is been developed, I'm certain	2-5-2	

ID	Text	# Need	Theme - Sub theme
	that it is robust, that the patient groups are relevant, and that it is likely to be better than current practice, it is got a high likelihood of that.		
11	Well, so a lot of papers who don't validate their results, right? So, paper who just do the derivation, but no validation. So, you don't really know whether that would be reproducible elsewhere.		
6	Is it strictly for my field or can it be generalized? If it can be generalized, then what are the things that apply to this generalization, right?		
19	I think that it is only the very large studies across multiple centers where you can interpret your results as saying clinically useful. Only those ones. You can't say it is clinically useful if it is just one center.		
20	The limitations and the usefulness of these is that they're mostly focused on physicians. And physicians like to think in terms of, they like to think in terms of decision trees or algorithms rather than in terms of probabilities, because probability is complicated. And, you know, it is not so clear what to do with a probability. It is much clearer what to do when you say, if you have more than six points, do this. If you have less than six points, don't do this.	2-6-1	Useful - Performance increase
20	You have to tell them what to do with the probability with the pooled cohort equation. For example, the American Heart Association, American College of Cardiology said run this equation and then if the probability of cardiovascular disease is greater than 7.5%, then prescribe a STAT. So it is very prescriptive about what to do. Because physicians aren't so good at sort of handling risk and understanding that 7% and 7.5% aren't really different. You know, or 7% and 8%. One of those above the threshold, one of those below the threshold, nothing magical happens when you cross that threshold.		
7	The stakeholders for me are the clinicians and the patients. So, when we talk about usefulness, those two parties should, there should be an impact there. If the information is purely statistical, then neither the clinician nor the patients will be really affected by those results.		
7	I would say useful results are results that can inform the usability of the prediction model. And in that sense, I would say it should inform on the degree to which the decision making in clinical practice would be affected.		
6	And it is good to have these discussion points in the paper as well. Like what is the potential of this solution that I'm proposing?		

ID	Text	# Need	Theme - Sub theme
11	So sometimes, for example, I've seen papers where, you know, the clinician prediction model would give, would just give a probability of disease, right? And then it says, okay, probability of, I don't know, like risk of bleeding in that patient is again, like 13%. Okay, but that sometimes doesn't help me in deciding, okay, does it mean that it is too high? And that should stop the medication? Or does it mean that, see what I mean?		
10	So in our case with asthma, it is a huge market. It is a huge interest and people are, the pharma's are, yes, yes, yes. We want this. We want people to think about modifiable risk of inflammation because we have nice drugs. So I guess that's another aspect where you need to think about how do I position my model strategically? So it is not only useful for patients and doctors, but also you get buy-in from a pharma.		
10	So three components to usefulness. Robustness, focusing on modifiable risks, something you can act upon as a doctor, not just as the patient. And third, having a nice way of integrating it in clinic… ,, want to focus on is modifiable risk. For example, in osteoporosis, bone density, you can act upon with bone mineralization medications, be it by bisphosphonates or biologics.		
10	And none are useful because they're not yet focused on modifiable risk. So I think when this point about modifiable risk enters, that's when people, you catch people's attention and people are interested. And I'd add that, of course, there's an added value of using modifiable risk is you can actually entice pharma also to be interested because when you work on modifiable risk, you're talking about therapies often, non-medical, but also medical.		
8	What does it mean to read useful study results? I think useful study results are something that I can walk away with and have, I should be able to walk away from the paper and summarize that useful result in just a couple of sentences.	2-7-1	Useful - Quicker to complete task
19	In your text, you just summarize what's in the graph or highlight certain things and the graph shows more. And so what you're doing there is you're highlighting what you think are the most important in the text and the numbers and people can see that. Then in the graphs, some people will see it better. A difference or something.		
7	What I would like to see may be some sort of additional analysis or simulations that could illustrate the use of the prediction models in real life. So now we have the prediction models. If it is used in that way, then the impact on the patient outcome will be that. And that could be just a simulation or something. And one can just acknowledge the limitations of such a simulation. But at least I think it should be should be done in a way.	2-8-1	Useful - User-control

ID	Text	# Need	Theme - Sub theme
3	Let's say if someone is just communicating and providing the percentages, but there's no end total number. It is just abuse of percentage.	3-1-1	Trustworthy - Accuracy
3	Yet the visualization can also like sometimes expose those things like which may be ... missed or you. But then. It is basically. It is a mix of both. Like you have some time you find something in the text which is unusual.		
19	I didn't trust what they, how they were presenting the results, because they presented a table with the true positives, false negatives, true negatives, false positives, sensitivity, NPV, all these things. And it was quite obviously wrong, the sensitivity, because they would present a sensitivity of 90 something%, but they had no false negatives, which means it must be 100%. So, I see that sort of thing sometimes myself.		
8	I think some things that I have seen done more often are using error bars or something that indicates the degree of error or the degree of change in a visual representation of a chart or a graph to make it seem as if the change that you're seeing in a variable or the difference between groups or something is much larger than what it is.		
8	Well, I think that inaccurate visual representations of data can be very misleading. So that can be important. I haven't seen this done. So thankfully, I don't think this is a common practice.		
19	So, if I see something and I go, I can't see how to repeat the study, then how can I trust the results but they can't check them? That's the point. So, if I see the way it is been interpreted, sometimes I'm going, I think they've got the numbers okay, but they're not interpreting them correctly. So, I might trust how they've come up with the model or the numbers, but from a usability perspective, the way they're interpreting it, they're overconfident, for example, that it could be useful or is really better than something else because they're not interpreting it very well. So, there I wouldn't trust conclusions, but I might trust the methodology or the results. Does that make sense?		
17	And papers are normally well-written, no doubt about that. So yeah, it is easy to understand what's going on. And the point that I made about it being sometimes understandable is what I said before, is that they don't touch base on the limitations of the model that's posing. So I feel like they make some bold, generic statements that may not be true for that mathematical model. So I question that a lot in my head.		
19	Okay. Well, I'm a scientist and we're meant to be skeptical. Sure. Aren't we? Yeah. When I see poor statistics and poor interpretation, I find it hard to trust the rest of it. And that is possibly a		

ID	Text	# Need	Theme - Sub theme
	little unfair because I think people are better at collecting data and running a study than they are at the statistics or data science. Hmm. Particularly if it is clinicians.		
12	Okay, so I think probably my response to your question would be that the single thing that would help is if there was some explicit part of the discussion that was more than the, that where people didn't need to be defensive about their discussion of strengths and weaknesses. I think this is a problem in this field that because everyone's so obsessed with my study must be the best, they put the weaknesses in but don't really get to grips with it.		
11	But at the end of the day, if you know that the prediction ability is not so good, you know that there will be a lot of, it would be very borrowed, right? So even if in that specific sample, it might look good, you don't know whether to a specific patient, it would be a prediction that would be accurate that you can take a course of action based on that prediction.		
8	And again, discussing thoroughly the limitations of the design or questions that are not quite answered by the study.	3-2-1	Trustworthy - Objective
12	And so more openness and explicitness that we're not necessarily racing to deliver the best possible prediction. … And however good we think the model is, here are the strengths and weaknesses of its clinical application.		
17	first of all, shifts into, where did they get the data from? Are there any biases in the data? Did they clean it properly? Can I reproduce that pre-processing pipeline? What assumptions did they check for?		
8	Yeah, so the limitations should be thoroughly discussed, and they shouldn't overstate the meaning of the results.		
17	It is also important to say in what context it is statistically significant. So to me, understandability means that you understand the problem, the clinical problem that is being questioned in the context of the mathematical method that was being used and all its limitations that it comes with and assumptions that it comes with.		
17	They might have done some preliminary statistics before embedding this into the model. But in the discussion sections or in their interpretation of the results, they don't talk about some of the limitations of the method itself and how that influences the findings or the results. So in that context, I find it not understandable because it is not enough to say something is statistically significant.		

ID	Text	# Need	Theme - Sub theme
8	They heavily discuss the limitations of the research and are transparent about what those are. I think that trustworthy research is something that, you know, I shouldn't walk away from reading that paper thinking, wow, they really missed out on this important limitation.		
4	If you want to exactly reproduce the results presented in the paper, then you need the model building code that authors used, the training model development software. This is extremely complicated to share. So I'm not really, I'm not supporting making this mandatory. It is helpful if it is possible, but it is very oftentimes it is not feasible... People will stop developing because it is become impossibly burdensome... The other reason is IP, intellectual property. Some of these software technology involves, for example, a deep degree of parallelization.	3-2-2	
10	It means that it is been derived and validated in a good way, in a good way, usually using individual patient data, that you've identified risk factors that are acknowledged or reproducible in their importance, that you've published in a good journal.		
4	But the other part of this equation is enabling others to make predictions using your final clinical prognostic model. So that doesn't require any code. You can put up a website where people can submit their input data and get results back. That's a lot more doable, and it definitely increases trustworthiness of the results.		
7	No, I think maybe at the very beginning when I started research, I was more resistant to that. But now I think it is better for everyone to have the codes online. ... especially when the methods get more and more complex, ... And sometimes it is more helpful to me to look directly at the codes. So then I know what kind of analysis has been done.		
17	So reproducible, I feel like there's some great work that's happened and I want to take it up and build on top of it, right? Then I want to know that the work that is being talked about is 100% reproducible. Oftentimes this is not the case, right? Even if you rely on public data sets, you don't get the same results that they're talking about. And I think the reason is because there is changes in hyper parameters that we use and largely because there are changes in the way we pre-process our data and prep it for data modeling		
8	You know, also, I think something that authors can do to avoid kind of p-hacking, if you will, is open science framework. so that anyone can look up stuff about their study, that's good. Having the data available either through an open science framework or through contacting the corresponding author.		

ID	Text	# Need	Theme - Sub theme
7	To me, it means that the report is transparent and reproducible, I think. Like, if I had the data that they used and did the analysis that is described in the method section, then I would expect to obtain the same. So in that sense, I would trust that, like, if I trust the study, it means that I don't need to do it myself because I expect that what is reported is what I would have found.		
17	They have clearly outlined what they did in terms of pre-processing and the entire workflow so that it can be reproducible.		
7	I think that's a good initiative and maybe something that would supplement that, the following of the tripod statement would be maybe some data sharing agreements. Like, if we could have access to the actual data, then that would be even better. And also, I like the initiatives of now more researchers, which is to provide the codes, the programming codes on GitHub or other repository.		
2	Anything that tells me that the analyst was honest enough to explore all the possibilities or at least more possibilities in the data, not all, I mean you come to all, but to explore the data more in depth.	3-2-3	
3	When you're reading the results, there should be a reason for every year action like the why you are doing this. And then you are also providing the evidence like I excluded this. But then I did the sensitivity analysis and then I tested this doesn't really buy the results.		
17	So the section of papers that I read happened to be ones that were written by clinicians and therefore it was easy for me to relate to and understand what was done and why it was done.		
13	But they use it as an example to illustrate what they want to say the methodological side. There should be a consequence and you should clearly state your aim in advance, so that everyone is knowing why we are doing this. For example, if you have a new predictor and you want to find a new lab value and say, is this lab value, does it give additional information on outcome?		
12	I'm clear about why certain outcomes have been chosen above others. And I'm clear about the population in which the prognostic model was developed.		
8	I think when the authors do not clearly explain or kind of don't make clear why this type of analysis is being done, which part of that is from the introduction and not part of the results section		

ID	Text	# Need	Theme - Sub theme
18	They usually start with 100 predictors that are described in the methods and somehow in the results you only find a table with 10 predictors. So what happened with the 90? Were they eliminated by p-variate analysis or just I don't know what happened. I think the paper, or when you report the risk prediction model it is like a cascade or a story. So I started with 100 I eliminated these because of missing data or whatever I eliminated these because they were not associated with the outcome or whatever and then I eliminated these because		
8	Well, I think trustworthy study results are ones that provide all the necessary information about the data collection procedures, the participants, any manipulations to the data that were done. So if they check the assumptions and data needs to be transformed or changed, or maybe there's missing data imputations or something like that, being really clear about that is very important.		
2	Well, that is a problem because when people do work on a question like this, they want some kind of predictive model, right? They want significant result. But the problem is that if they don't get what they expected, then they start panically asking for more sophisticated models. And they are doing the thing that I just said, they're causing overinflation of false results. But sorry, I kind of drifted.		
12	So, I would say that that, you know, that I would be concerned that there is a clear there is a clear justification for things like the number of variables that are being used and so on that might be done in purely statistical mathematical terms that, ... I know these are contested areas, but I think still that needs to be explicit, because again it is this problem of just reading another paper where they've used a data set, plugged it in, and out comes the evidence at the end.		
14	And then the question becomes, should we even be doing this? Should we even be doing this? Should we be building models and publishing them and acting on them if the sample sizes were probably too small? ... It is better to have some evidence-based information, make these decisions. And then some people would say, no, that's really kind of dangerous because having too small of a sample size probably would lead to unreliable or unreproducible results. And so that's kind of the, that's what I struggle with every day. Should I even be doing this if I think my sample is too small?		
18	What pieces of information do you look for to make your decision of using model A or model B? And once you answer that for yourself we should try to always report that.		

ID	Text	# Need	Theme - Sub theme
17	That all the assumptions are checked before using a mathematical approach, that the results, the sample size is big enough for that particular method so that you don't have that curse of dimensionality problem.		
8	So I think tables can be good. I think that if there is missing data, being really clear about those missing data procedures. And I think even you can include information about missing data in like a participant demographics table or something		
3	That's a study has only one thousand data, but there is only 50 events of like the outcome and then the reader just follow some standard scripts and maybe author has just follow those standard scripts and replicate the results. You don't really feel confident. You feel like the how much investment has been done, because if you haven't really justify your sample size, it means that you're not really serious. So you can see like that.		
13	Maybe just to say one topic which is always a topic in discussion is how to treat missing values. Maybe just to say one topic which is always a topic in discussion is how to treat missing values. Even me, I am learning. I try to replace everything that was missing. Sometimes it is not a good idea. Then I had a model where every missing is accepted. It would not change the result if something is missing. I think a mixture of both is maybe the best.		
1	I think that you want to know that I guess missing data has been dealt with nonlinear effects. And I think that I find usually they're sort of trustworthy in that sense.		
19	I want to see that, particularly around that primary outcome or primary analysis, that it is not oversimplified, that they have looked at it from more than one angle. And those results are presented And those results are presented There's more than one simple analysis. Like if all I'm producing is AUC, and nothing else, it is insufficient. So, it is the same sort of answer as before.	3-2-4	
6	I think, at a minute, like you should show how you do compared to that model ... but ideally there's a, an actual like a reasonable model for baseline,		
6	I don't think it is done very often but I'm curious, would be curious is if there's an area where there's kind of a baseline prediction model or an already used tool, like how does this model compared to the baseline.		
6	Yeah, but it is ideally, you should be showing not just that you do a good job you should be showing that we do like a better job, or we do a comparable job but our model is simpler, you know, which I don't think is the norm. From my experience reading literature.		

ID	Text	# Need	Theme - Sub theme
6	I'm sure that somebody out there is working on a better cardiovascular risk prediction model and I know that there's a, you know, coronary heart disease risk prediction model. If you come up with a new one, you should compare it to the one everybody is using and not just present your results on their own.		
2	Cross validation or any kind of simulation modeling might provide another layer of robustness. I think that the authors can do some small tweaks like these. Maybe they could explore the subset by subset and then try to explain more about mechanics in various subsets of data. So basically, anything that shows me some kind of beauty and humble approach to the data would make me think of the results more trustworthy. If I had only seen the results of the model. We have the model, this is the model and this is how it is, I'd probably be substantially off put by this and I wouldn't treat it as something trustworthy.		
2	The basic one and the easiest one would be some kind of cross validation, where you run the random subset or 70-30 or whatnot, one against the other, and you see the model results. Of course, the main question is whether the statistical power remains the same. Well, if the cross validation fails, then your overall sample size was too small to begin with, and then immediately your results are less reliable than we would want them to be.		
2	I would guess that most important thing would be if the analysis side has sufficient power or skill to fix an effect, and then see what happens with the model if this effect is fixed, if this is removed, or if you have small sample size, then do the leave one out. What happens with the model if you drop any kind of patient from the model, where you see how the patient or a variable fits the model.		
3	So of course, results are not 100 percent perfect .. So that's why I said that when I read the paper, I feel like the how much efforts are being done and then how can like they can improve those results by using some like either the tripod or there is some other guidelines in terms of assessing the biasness of the studies.	3-3-1	Trustworthy - Validity
10	Well, I mean, not reporting obvious important elements as dictated by the checklist. Otherwise, as you see in the law, no, I haven't. I haven't seen sticking a lot of things in supplements. Oh, omitting voluntarily omitting important risk factors or important demographic characteristics.		
11	Well, that's where for me to trust it, I need to make sure that the whole derivation adhere to standards of derivation validation. Sometimes it is too often like a convenient sample. People took whatever database they used, and then they use variables that are in there.		

ID	Text	# Need	Theme - Sub theme
9	So, I mean, I just go by the standard just researcher checklist of, you know, do I trust this? Do I trust this article? Do I trust these authors? You know?		
18	It is not that I think they are lying or making up things or fake data. If you don't find all the information that you are looking for or that you will expect to be basic information to be reported We cannot really trust in that model.		
1	They feel like they're more interested in getting the paper published than actually developing a prognostic model, and the lack of experience just is immediately obvious. They won't have followed the guidance properly, they might say they followed the guidance, but you can tell that they haven't. Then it is very difficult to review, because as I said at the beginning, it is like to review it properly, you have to give them so much feedback, you almost then warrant becoming co-author, and it takes a lot of time.		
14	And so there are, there are models that we've been making using to make decisions for kids for 30 years that have never been formally validated stuff ends up in our literature, and we act on it. And we use it, but it is never been validated. It is like I think of a survey instrument, people say, Oh, that's been validated, but they don't understand that validation is specific to the population, right so this they say this has been validated. Well, it is been, you know, used in a group of patients who are that are different than the patients that you're using it. So I think that people don't understand what validation really means or what it, what it entails.	3-3-2	
13	I sometimes saw papers when they applied the scoring system and did not consider that the population was not according to the original one. This is what I saw. The authors that publish a prediction to the usually describe the population is a very interesting example of a research where some people apply the new score to their patients and publish the results. There are some people that do this. For example, they included a lot of patients with minor injuries where the score was not made for. This is severely injured and in need of intensive care. This is a bit like the study results. It does not work.		
11	Clearly, I find it when you see sometimes that it is kind of quick and dirty, right? That, oh, we took that and there is no attention to all these, right? And the validation one is a big obviously, if they did a just a derivation and just one logistic regression, but they don't, they didn't replicate their findings or they didn't, or so that's when I certainly would not trust the result.		

ID	Text	# Need	Theme - Sub theme
13	We use data from three years for development and next year for validation. Others make random selection of some cases for validation. 75% were used for a development data set and 25% for a validation data set. These kinds of validations usually work very well the more likely you are to apply this. The more likely you are to apply this. The more likely you are to apply this. If the sample size is sufficient let's say 1000, 2000, 5000 there is no problem. Internal validation will work always. It is very interesting if it works with external validation.		
2	Replication. Period. Ideally replication in unlinked sample.		
2	Cross validation or any kind of simulation modeling might provide another layer of robustness. I think that the authors can do some small tweaks like these. Maybe they could explore the subset by subset and then try to explain more about mechanics in various subsets of data. So basically, anything that shows me some kind of beauty and humble approach to the data would make me think of the results more trustworthy. If I had only seen the results of the model. We have the model, this is the model and this is how it is, I'd probably be substantially off put by this and I wouldn't treat it as something trustworthy.		
14	And yeah, so, I mean, the methods need to be aligned with what the goals of the study are. ... So being very clear about that and having the methods meet the goals. I don't know if this is exactly addressing what you're trying to get at.	3-3-3	
19	So, for me, the confidence comes actually when I first start reading the methodology. If I can see that the methodology about how they've done things, how they've chosen the patients or collected the data and done the analysis is robust, that gives, and repeatable. And the point of the method is it must be repeatable.		
4	It basically comes down to reviewing if the authors used sensible methodology in their research. And if they have clearly explained it, and there are no gaps or obvious holes in that. I guess it is not that difficult to find those gaps. If there are some.		
18	I think it all goes back to the methods section of your paper or thesis or whatever report you are working on. That's where you should really gain the confidence of your reader by reporting everything clearly and transparently. I think it is sometimes very difficult but we have to put ourselves in the shoes of the reader or even better in the shoes of those who will eventually use the risk prediction model and ask ourselves if I were to use this model, what information do I need? And obviously report that information.		

ID	Text	# Need	Theme - Sub theme
8	So I think, you know, first, like in the methods section, being as precise and clear as they can about the data collection procedures, in the data analysis section, being as precise and clear as they can about the analytic procedures that were done, as well as any manipulations to the data, any statistically informed or theoretically informed choices in the statistical models, for example, if we're going to use covariates. What covariates are we using? And provide citations or clear justification from the current study data for using those covariates.		
8	I have not seen examples of this, but I do think that possibly even having a table where they go step by step in order of the analytic procedures that were done. Like maybe first, the data were checked for normality and skewness. And then second, the data were log transformed. And then third, whatever, maybe there were imputations done or something.		
4	The other is actually predicting, generating predictions for new data. And this is completely different things.	3-3-4	
2	So I guess replication is the key because if the results are replicable, then it means that I can apply them to my situation. If they are not replicable, then you are in a whole field of possibilities what might have gone wrong. And you don't want to be in that field because it can be the data, the context, the model, the assumptions, the parameters, anything can make a model miss.		
9	Oh, I mean, for me, because I'm a researcher, I mean, it is just like the standard that we all kind of look for. I mean, whatever, like if I look at an article, if it is trustworthy, I mean, can you replicate the results? You know, that sort of thing.		
2	Ideally, independent data set replication, and then you know you have the real deal. Otherwise, the models are so diverse that it is almost difficult to say you have to do numerical or you have to do visual or you have to have this as measure. I guess I'd like to see,		
12	Now, we've had a series of really excellent papers from America, especially and elsewhere in Europe, that have tried to make the prognostic model better, and more precise in its prediction, and where they haven't been able to replicate the results, that if you apply it in practice, that people do better if the model is used, and if the model is not used, which is the ultimate thing, isn't it, for prognostic models? Do people, do our outcomes better, if you use it to determine and guide treatment?		
10	Yeah, well, there's the checklist. There's a checklist for clinical prediction models. I forget what is it? Is it? Yeah, I had strobe in ..., but yeah, ... or whatever. Yeah, so that's helpful. Of course,	3-3-5	

ID	Text	# Need	Theme - Sub theme
	it doesn't apply universally, but at least it provides a checklist where you can say, well, this, I have this, I have this, I might as well mention it. So I think		
1	Well, I don't think they help either. Calibration, discrimination. I think they're too complicated. There was another plot that I think it is mentioned in tripod and I mentioned it. Decision curves. I think ... I'm not sure that the tripod goes far enough in giving people the right guidance for interpretation. Because I think probably the tripod authors understand those plots, but fail to recognize that the average population don't understand those plots.		
3	So it varies really. So how this paper is actually what is a workflow or your pipeline of this paper is, are you this study is how this came along. So at the end product is a lot of factors. It is not really the one factor that author choose, but it is all over the how this has been reviewed, how the journal they are considered for this. So I know that the things are like the mixing a lot of things, but this is just for the reporting the results, the tripod is just, I think, the tip of the iceberg. It is just, it is not all things.		
7	I would say I saw the reverse where in my sense the methods were more attractive than the actual use of the paper at the end. Like, for example, the paper in Nature that I mentioned earlier. The method was great, that was very modern methods, deep learning, very trendy AI techniques, but at the end, in my opinion, the clinical prediction model was not useful because the sample selection was not appropriate to address the research question.	3-3-6	
13	First of all, we have the point that did the authors have selected the right population? Or is it some kind of artificial subgroup? You think about future application and the group of patients a trustworthy study result? What does it mean? The population used to calculate the prediction model should fit to the population where you think you could apply this. The population should fit. The population should not be a small population. The larger the population, the more likely you are to apply this. The more likely you are to apply this. You have a model and apply it to new cases and then see whether it works.		
14	And so, if you, if you're going to be publishing a model, and the stakes were really high, then I would want to make sure that, you know, that I trusted the entire process of the derivation model that the population was representative of population that we're talking about that there were large numbers of patients that there were that these variables are very high.		
14	So, they would meet, whatever this model was for it to be trustworthy, it would have to meet certain criteria that I have. And I have to admit that I have published model that I don't know		

ID	Text	# Need	Theme - Sub theme
	actually meets this goal of trustworthiness, only because our, the population, the sample of patients I had to include in this model is too small. So, performance metrics look good. But we know that I, you know, I probably needed 2000 patients instead of 200 patients so that makes me a little nervous.		
18	It probably or most likely was developed with the best high quality data and with a very good population but if you do not find the information in the paper you do not have any way of actually knowing that.		
10	And fourth, I wouldn't insist, trustworthiness is very important, but I wouldn't, I think I put a citation in there, all models are wrong, but some are useful. Yeah, you know that citation? Yeah, yeah. So I think that's, again, you need to have a model that's trustworthy, but even if it doesn't pan out perfectly in practice or in another prospective external validation cohort, if it is useful, if the data's trustworthy because it is based on clinical trial data, I mean, that's helpful.	3-3-7	
20	I don't know anything about how good the data are that go into your model, right? You tell me that you identify these outcomes. I don't know whether those outcomes are validated or not. Most papers don't say how they validated the outcome. And then you have a bunch of predictor variables, and I don't know how you validated your predictor variables. Most papers don't say.		
11	So if, if there is a code for hypertension, how sensitive and how specific that code is at truly detecting the hypertensive patients. So you may be like a color code or, you know, I don't know. And then, and the same about, is there a clear definition, international, you know, standardized definition for that variable or not?		
11	Like, I like, you're trying to find ways to, to, yeah, to better illustrate things, right? Well, maybe, maybe a table or a figure with all the key variables and showing their level of validations and lab.		
12	I think a poorly done thing in a lot of medical record studies is an explicit definition of the levels of a variable in the medical records, so that, you know, understanding exactly what the coded diseases and disease labels mean and incorporate and stand for. And once again I think the problem at the moment is that that's my traditional background saying that you must understand clinically what this group of conditions are, what you're putting into them, what the doctors that are involved in recording or the nurses, you know, is this a very blurry thing or is this very precise with international consensus over how it is done.		
11	maybe, you know, again, if it is from an EMR, how accurate the code is in that database, right?		

ID	Text	# Need	Theme - Sub theme
11	Or obviously if these are new variables, or if they were never used in predicting a certain outcome, I would expect to see some data about definitions, standardization, reproducibility of the variables as well, just to make sure that the different physician would interpret that the same way and scores that the same way.		
11	But it is very rare that people would really start by getting a systematic review of all the potential predictors, then collecting the data in a specific big study aimed at deriving the score. So if I see that, oh yeah, we use data, have a label wherever, and we use the variables that were there, I'm less, I would trust that less.		
8	Also, of course, I do believe that it is standard, but I do think this is also, it is not revolutionary, but I think part of what makes data trustworthy is when the authors are providing the internal consistency values for their own measures and things like that.		
12	And, and, and that the output in terms of, for example, the absolute, the absolute risk that's being predicted, and these sorts of things are clear, and they're appropriate to the outcome that's being measured, you know, is it a, is it a, an incidence rate or, or what it is?		
7	Of course, it should be made within the limits of what is possible, given the data. But at least that's something that one needs to keep in mind, because we often think a lot about physician medicine, how to treat the individual. But then the next question is, when we apply that strategy to everyone, do we actually do something good to the population?	3-4-1	Trustworthy - Stability
7	If I have a prediction model that allow me to predict, I don't know, a patient outcome, and I will have some patients with different ethnicities or racial background. If I take them one at a time, and I make a prediction and I say, oh, this patient's risk of probability of death is this high or that high, I'll treat the patient according to the prediction model. Maybe I have the illusion that I'm doing good because I have just that one patient.		
7	It was not in health, it was about criminality, and that was a high probability of false positives in black people compared to white people. Here, that's kind of the same issues that I think we need to be aware of when we work with clinical prediction models, is that do we create health inequality across different strata of the population due to the skewed performance of the prediction models? If prediction models work better in white people and work very bad in another group, then it might create inequalities.		
7	But then another question is whether the systematic use of some prediction models actually help at the population level. Like when we just look at the individual level, sometimes we have the		

ID	Text	# Need	Theme - Sub theme
	sense that we are helping, but at the population level, we are actually creating more inequality or things like that.		
12	So we've got to improve our prognostic models. And so all the time, we want to reduce the variability, don't we? But always, we will be leaping off at the point that saying, this will benefit this subgroup. But we don't know which, you know, necessarily which people. So population and individual is quite an important and subgroup and individual is important.		
12	So there's this sense that the algorithm that functions very well for the average, and for the population, and broadly speaking, gets people in the right groups. And for broad, urgent policies, as we needed in COVID, that's okay. But that might not work for the individual sitting in front of you, who might have a condition that doesn't appear in the model, and which you may need, you know, you may need to think round how you're going to deal with that.		

APPENDIX H

The complete list of the paragraph chunk included in the qualitative analysis to answer RQ1 Chapter 6

ID	Text	Preferences in using visualizations
5	Actually, figures to make it more communicable to readers. For example, for studies of COVID-19, how to calculate the model and apply the results. Those figures are necessary.	Using visualizations
3	And for the modeling, I think the communicating the result better, the visualization is a core as well. For publishing the paper, table one is important, but for the like the communicating the results well that the visualization is a better one.	
17	And then also maybe having some mandatory figures that if the journal mandates that I need to look at. ... So with this, if they made a section in the results for explainability or understandability, I think that would improve the quality of our papers a whole lot more.	
8	But doing things like changing the dimensions of a chart or a graph to make it seem as if the change that you're seeing in a variable or the difference between groups or something is much larger than what it is	
10	But even better when there's colors, different layers of information, where the quantitative display of information is rich and you can keep your eyes on the figure for two minutes and always learn more, and where the figure speaks for itself. So you don't necessarily need to read the article, but you see a nice figure, you understand immediately what it means. So figure table.	
19	I find when I talk with some clinicians and maybe just some people, some people are okay with numbers. Some people, they need the graphs.	
17	I think one of the reasons is that there is a limit on figures that one can use. And so oftentimes, even if researchers add this, this just gets into supplemental sections or something. So maybe increasing the number of figures that we can use is one option to promote more of this.	
3	I think the visualization is the core part of the modeling. That's your first step that you do. That's definitely is the… Yeah, when you are presenting the results, that the visualization is the	
3	I think the visualization is the core part of the modeling. That's your first step that you do. That's definitely is the… Yeah, when you are presenting the results, that the visualization is the… Of course, it is a… As a person, as a practitioner of the… Like maybe that you actually better understand the visualization is better than the text. But of course, some people follow the, like, as an appendix or a supplementary material, you provide more text and more narrative around the, like, what the meaning of this craft is.	

ID	Text	Preferences in using visualizations
7	I think visualization would be nice, like graphs or figures that show the potential effects of clinical prediction.	
12	I'm a traditionalist, and so I like to see the graphs and, and that's helpful. And I think, I think discrimination, for example, is an interesting one for understandability.	
18	I'm sort of in favor of that opinion. I don't know if it is to make them more understandable, because I think the basic figures can make them understandable, but I do believe that fancy figures or new visualizations, I don't know if they will make it more understandable, but perhaps more attractive or more appealing for an audience.	
18	I'm sure if you ever get a journal, a physical journal, and you just go through the pages, you're most likely to stop in the paper that has the most nice features. So again, simple, I don't know, rough curve is not very attractive.	
6	Maybe too many tags too many visualizations too complex visualization, that sort of things. I don't know if I've seen too many visualizations. I feel like I'm usually left wanting more. I feel like the statistician or whatever is probably the wrong audience for that now mind you, most of that stuff belongs in an appendix, you know, there's certainly you could have a poorly written or poorly flowing article by having too many figures in the main text but that is different, you know…	
18	Maybe you have some way of putting colors or making it more fancy, maybe combine different information in one figure or in different panels. I do think that will seriously boost your chances of getting into better journals or just getting people more engaged with your paper.	
17	Oh, a whole lot. I think it is very important to have figures and tables that consolidate what the main findings are. And I think people generally do a good job of this.	
2	So again, visualization, excellent for understanding. I guess maybe I've been a bit harsh. With proper, well, it depends on the data, with proper method and good difference visualization can tell you whether something is similar or different, but still, I'd want to see numerical output as the main result.	
3	So as you know, the images are the pictures are more informative than the text. And using this tool, the results can be better explained. So I would be inclined to the visualization more.	
2	So I guess that visualization is fine. It is excellent for the first step for you to kind of think in which direction to go. But definitely I would not like to see the paper that is only driven by visualization. I would want to see some decent statistics in the background. Here I'm conflicting with new papers based on mainly AI propelled thingies, but these are even worse in terms of standardization. So I guess that my own viewpoint is that visualization is excellent in the first stage. It can be useful for the entire study, but I would	

ID	Text	Preferences in using visualizations
	like to see numerical output with formal tests just to be sure that it really works.	
10	So I think you can use different ways. So like, are you suggesting multiple figures for, let's say, one specific models to help? I think multiple ways of applying it. So I think it is not heresy to have the regression-based equation, but then you show a figure and say, this is what it looks like.	
10	So I think you can use different ways. So like,I think multiple ways of applying it. So I think it is not heresy to have the regression-based equation, but then you show a figure and say, this is what it looks like.	
10	The first point is the figure or the table. I think you need a good figure or table. It could be a forest plot where you see quite clearly the risk factors.	
3	The most of the time this based on the visualization. But of course, it is needs a narrative as well. So it is like the for example, the same kind of I have I have seen recently very funny decision curve analysis. People have developed the model that they like that they're using the standard packages. They develop the graph. But nobody knows what's going in the graph. And the author doesn't know.	
13	Yeah. So I will prepare tables or figures, if I think they will be helpful. And I show it to the doctor, and he said, oh, very nice, we take it in the paper.	
7	Yes, less understandable. First, when it is only text.	
3	Yet the visualization can also like sometimes expose those things like which may be missed. It is a mix of both. Like you have some time you find something in the text which is unusual.	
8	You know, again, I think it depends on the complexity of the results. I think if the results are very complex, having multiple visuals can be helpful, rather than putting all of the relationships into a single visual model. Sometimes that can be helpful, especially when the results are very nuanced.	
8	I do think that, you know, you've been asking a lot about visuals. Personally, I like to have tables and figures that are referenced in the text because I think that makes it much more parsimonious to like read the results and kind of move along.	Accompanying visualizations with text
18	But of course, we shouldn't expect everyone, all the readers to understand those figures if that's not their field of research or expertise. That's our sort of obligation as authors to put that into context and clearly explain what we're showing.	
17	I think the most important piece is that the text that you're writing, for me as a reader, what I see visually needs to be set in context in the text as well. So it is nice to have that narrative of, oh, look at this figure two or table three or figure four. Here's what we mean. And have that indicated either in the captions or in the text.	

ID	Text	Preferences in using visualizations
7	I think, I think both. Both. I think the visualization should, how to say, visualization should allow us to imagine the product and its value. But of course it is hard to summarize everything into just one graph or one figure. So, in that sense, the text is always nice to supplement things. But the text should not be the first thing to look at, because it is often very heavy.	
19	I was always taught not to put numbers down if you've got it in the graph or graphs down when you've got it in the numbers, so you don't double up. But I think there's actually some value in doing that. Whereas in your text, you just summarize what's in the graph or highlight certain things and the graph shows more. And so what you're doing there is you're highlighting what you think are the most important in the text and the numbers and people can see that. Then in the graphs, some people will see it better. A difference or something.	
20	Or like what kind of balance there? Yeah, I mean, I think visuals are very helpful, especially when you're talking about a continuous predictor. But I don't know how much, you know, how much text versus how much visual. I think it would just depend on the, you know, how well the author communicates, you know, in words versus in pictures. Yeah.	
17	So if the reader is not guided on what it is that they're looking at, it is one thing to have a bar plot of features and say, oh, this is important. But if the reader is not guided as to what they're looking at or in terms of, let's say, in terms of a pie chart or a statistical significance display, if in the text you say, oh, this is largely important because, oh, when you go into an entire clinical spiel but your figures have a small silver pie piece on the chart, then that's questionable.	
8	Yeah, I have seen papers where results are only explained textually. And sometimes that's okay if it is a simple research question with a simple statistical analysis. But I find that often, it is not helpful to only have a textual explanation because it is like you're just keeping a list in your head.	
8	I think one example of parsimony is, you know, if you're going to use a visual, have a minimal, a concise and clear textual explanation of the results without discussing too much or too little.	
8	Yeah. I think that sometimes too much text makes it complicated. I think having, if the authors are able to create a good visual, then that should be paired with a minimal textual explanation.	

APPENDIX I

The complete list of the paragraph chunk included in the qualitative analysis to identify visualization tasks in Chapter 6 (RQ 2)

Section	Visualization tasks	ID	Text	# Need	Theme - Sub theme
Data exploration	Detecting initial data patterns, trends, or anomalies, outlier	17	I think this is very important, especially in the data discovery phase. First of all, to show the readers what signals the data may have in the beginning, just to show what kind of characteristics the data has, how much variability there is there in each of the features that's being questioned. Are there any ceiling effects or floor effects in the outcome variables that we are interested in? Those kind of things, just to first of all, present the data. Data visualization can go a long way. And again, this is not so much explored in the results section of papers.	1-3-1	Understandable - Analysis
		14	you have to be able to explain that other things going on behind the model of variable that seems like it shouldn't behave the way it is makes it behave that way in terms of, you know, how much variability there was in it relative to other other variables in the model or whatever it is, you know, it is even hard for the statistician.	1-3-4	Understandable - Analysis
		19	In table one, where you're just describing the demographics, or other information, if they present, say a blood measurement, and they present a mean of 100, and a standard deviation of 300, I don't trust that, because I know it is not normally distributed from that. If you take one or two standard deviations from the mean, you get a negative concentration of your blood, which is nonsense. So, they don't really understand, you need to present a median, and	3-1-1	Trustworthy - Accuracy

Section	Visualization tasks	ID	Text	# Need	Theme - Sub theme
			interquartile range instead, or maximum, minimum as well, or do it graphically, so you can see the distribution.		
		3	But the visualization could be a you can see find some nonlinearities or some kind of unusual outliers in visualization. But then nobody nobody really comments on that. Like why there is this outliers and what is the consequences if you exclude them or what if you add them. So these kind of debates can be initiated from both sides.	3-2-3	Trustworthy - Objective
		17	So in terms of authoring this, I think it is very important to show all the groundwork that we do in terms of data cleaning, data pre-processing and checking the data for all the assumptions and then fixing, like there are some, you can do a test for normality, for example, and then you can apply some transformations to make sure that your data is correct. It is difficult to author this because like I said, word limit and figure limit make this very unappealing and most of this just gets summarized in one single sentence or two sentences at best.	3-2-3	Trustworthy - Objective
Predictor selection	Visualizing predictor selection	18	Yeah, I mean, I think that from the reporting standpoint, it is just very simple to have a table with all your coefficients and intercepts. It is not that complicated. And it shouldn't be like, you shouldn't go into too much trouble to have a fancy figure for that. It may be as simple as a table with all coefficients. And also, sometimes it is not even clear, like, what the final predictors were, because many,	1-3-2	Understandable - Analysis

Section	Visualization tasks	ID	Text	# Need	Theme - Sub theme
			many times they start with, I don't know, 100 predictors and then they skip those that were significant in the deep array dissociation and so on and so on. And it is that those steps are sometimes not even clear so you might read the whole paper and you didn't actually learn how or which the final predictors were. Anyway, but you can have a simple table with these are the predictors, these are the coefficients and this is the intercept. That's not that difficult.		
		18	They usually start with 100 predictors that are described in the methods and somehow in the results you only find a table with 10 predictors. So what happened with the 90? Were they eliminated by p-variate analysis or just I don't know what happened. I think the paper, or when you report the risk prediction model it is like a cascade or a story. So I started with 100 I eliminated these because of missing data or whatever I eliminated these because they were not associated with the outcome or whatever and then I eliminated these because	1-3-2	Understandable - Analysis
		17	Sometimes, this may not be the case that is relevant in terms of how it really helped the model. For example, if it is a case of linear regression, I find most clinical papers do not go into issues with multicollinearity, which is a very important aspect about linear regression. You want to have all your features being independent.	3-2-3	Trustworthy - Objective
		17	But beyond that, I feel like as a field, we don't go beyond to try to come up with data visualizations that show more properties of the data that helped us build those predictive models.	3-2-3	Trustworthy - Objective

Section	Visualization tasks	ID	Text	# Need	Theme - Sub theme
Modeling	Explaining model architecture	17	But I think if it comes back in the AI machine learning world, we can show them in graphical representations or visual representations as well.	1-2-2	Understandable - Comprehension
		17	And the impact that has on model development as well. So we can, I think it is important, we can use visual representation to our machine learning models not only in how they were designed and built, but also in the interpretation of the results and findings, I think visualization can come a long way.	1-2-2	Understandable - Comprehension
		6	Most of the time, it can hide. It you know it is not easy to tell where the actual, like, say black box model and the simple representation differ. You know, it is tricky but I mean, at the same time, nobody's showing what all you know, it is not really possible to show what all possible that outside of like trivial prediction models.	1-2-2	Understandable - Comprehension
		12	But maybe more of an issue around explaining components of the methodology and how variables were selected or in particular and increasingly these days what goes in the black box of machine learning or artificial intelligence type tools, which is the particular problem, ... That's where I would want quite explicit criteria of either understandability or the researchers owning up and being clear that actually there's a point if you're doing a black box type of neural network or so on that that might not be directly understandable and we have to take it on faith.	1-2-2	Understandable - Comprehension
Result exploration	Highlighting predictor importance	13	And another option is if it is not that primary goal to have an exact prediction model, but have a model to say which factor has which importance. So for	1-3-4	Understandable - Analysis

Section	Visualization tasks	ID	Text	# Need	Theme - Sub theme
			example, you have like 10 different predictors and you want to see what is the ranking, what is the most important one, and so on and so on.		
		6	An alternative I guess you could sort of do what I mentioned, you could say, okay, I don't have model coefficients, but I can tell you how much what the AUC decreases if you were to remove this information from the model. So, those are just something that lets you know these are important, and also try and quantify how important they are.	2-1-1	Useful - Critical support
		6	Yeah, so I think having kind of a top five or top three, or something like that for maybe both harmful and protective effects, or, you know, maybe just top five by magnitude and say whether they're harmful or protective. I think that is a good way to maybe present at least some of your results and hopefully, you know, hopefully there's something there that is immediately obvious and like you want to make sure that you capture some, some of what we already know, but the hope is maybe you're also adding something that is new, or at least isn't as solidified.	2-1-1	Useful - Critical support
		6	Yeah, so I think maybe one way you can visualize it is if you have all of your predictors. You could rank, you could plot. How much the say AUC or MSC or some performance metric. And if you were to remove that information or randomly scramble it. So, sort of like along the x axis you have all the predictors and then you've got, here's the, the change in performance. So, you can see the change in the predictor and how it affects the performance. .. I think everything I have, I	2-8-1	Useful - User-control

Section	Visualization tasks	ID	Text	# Need	Theme - Sub theme
			say has to come with an asterisk you know like it depends on your model.		
		6	So, if you have a model that has coefficients you can, you know, show what those coefficients are, you know, like if it were just a simple linear model then you can say here's the top five most important predictors by the magnitude of their effects and say what that is, but not necessarily always possible, depending on the model that you picked.	2-5-1	Useful - Useful for tasks
		6	You could say, if I didn't have age information. So you fit the model without age. How does that perform. And then you could just for like visualization purposes, I would probably sort that so that you have either increasing or decreasing importance. So, like, what's it called a lollipop chart or something like that where, yeah.	2-6-1	Useful - Performance increase
		6	but you know if you say that you only want to have like a maximum of three predictors in your final model, and it has to be some sort of score, you know, like integer scores coefficients. If you change some of the input data, because some predictors, you know, may do as so like I guess. You know, using weight versus BMI,	1-3-4	Understandable - Analysis
		10	So I guess first, the quality of the writing. So some people just write better than others, I think. Actually, that would be the second point. The first point is the figure or the table. I think you need a good figure or table. It could be a forest plot where you see quite clearly the risk factors.	1-3-4	Understandable - Analysis
	Demonstrating relationships of	13	So for me, it is very unsatisfactory not to know how the work in between goes. So the models that we	2-5-1	Useful - Useful for task

Section	Visualization tasks	ID	Text	# Need	Theme - Sub theme
	predictors and their impacts on model output		create, we list the influence factors and we show the effect of each factor.		
		6	I ran into this problem when writing my own paper but I don't think that there's a great way because of visualizing some of these models, like you can some things you can do, or some visualizations I've seen that like kind of help is, you can kind of fix a bunch of the other, you know, fix the other predictors that their main values or something like that. And then take, you know, your most important risk predictor and show over the range of the possible values what the change, what the risk looks like to help visualize.	1-3-4	Understandable - Analysis
		6	Yeah, I think it is kind of an open research question, because also just as kind of a follow up to what I was saying you can kind of, instead of holding everything constant and letting one covariate vary, you know, is there an intelligent way of, you know, we want to see what happens with age across the lifespan, can we let the other covariates vary in a way that is sort of most representative at that age group. I don't know if you were. It'd be tricky.	1-3-4	Understandable - Analysis
		6	You of course run into the issue that sort of just not, it is not the same issue but it is related to what I was talking about before, where you keep your holding everything else constant and changing something so you can end up in parts of the, you know, you can end up with combinations that never really appear in the data, potentially, if you hold everything fixed at a certain value and then vary age, you know, because 18 year olds probably have different, you know, number of comorbidities or something like that, than	1-3-4	Understandable - Analysis

Section	Visualization tasks	ID	Text	# Need	Theme - Sub theme
			90 year olds. And so, by holding everything constant, you know, it is still useful to see what the shape looks like, but, you know, I'm trying to think what else has been so you're, we're still talking about understandability, yes, the model not, you know, performance necessary so not like an ROC curve.		
		8	Yeah, I think that if you have variables, if you have a lot of variables that have very complex relationships, visualizations can be super helpful. If you're doing a path analysis model or some kind of structural equation model or something like that, that can be very helpful. Especially to see, I like to see both a predicted figure, like if you're using a figure of a predicted model, I want to see what we're predicting is going to happen and then see a visual of the model of what actually happened, representing which relationships were significant and which were not, what directions they went in. But it depends on the models that the author chose.	1-3-4	Understandable - Analysis
		19	how the hazard ratio changes with that variable or the odds ratio. And you can do that using restricted cubic splines to model the variable. That is a problem when people split things up like that and just try and present things, is it is never how the, it is just not realistic. And you throw a lot of information away when you dichotomize your variables or your predictions, you throw a lot of information away. So graphically it is often, you can show that much more.	1-3-4	Understandable - Analysis

Section	Visualization tasks	ID	Text	# Need	Theme - Sub theme
		3	And the one thing that came in the mind, like if this is a continuous outcome, you can use a scatter plot or something like that. You can see the relationship. But when it is a binary, how you do that? So in this case, like the visualization is really important. Like if without visualization, you actually can't really see. And then you have to be very innovative. There's no standard plot that can actually show this relationship. So you have to be innovative in a sense to see the relationship of those variables and which is really good in terms of communicating the results to the non-students. Those who are not from your field, they can actually understand better.	1-3-4	Understandable - Analysis
		2	I'd like to see some maybe even simulation like drop one out or something from that category for us to see how robust the results were. If that fails, then already you're kind of in trouble, because even the high R squared or S value do not necessarily mean that the model makes sense, because they are all just numbers. And again, you can maybe fiddle around and get them to some kind of reasonable level. And it can look good. But in the end, it doesn't mean it works.	1-3-4	Understandable - Analysis
		6	Say like a model with one predictor does this well or 10 predictors does this well and make a curve like that. I was also thinking that you could say, if you could specify the predictors so it is like the x axis is not really. It is just categorical, you know it is saying like age, whatever.	2-8-1	Useful - User-control
		2	Ideally, with modeling, you would want to see some kind of almost even a simulation where the expert then explains if you change some of the entry	2-8-1	Useful - User-control

Section	Visualization tasks	ID	Text	# Need	Theme - Sub theme
			parameters or model assumptions, then you get results like this. This means that this factor is important, other ones are not.		
		17	It is important for me for results to be useful that they're not just statistically relevant, they're also clinically relevant. So instead of just stopping at region X, Y, and Z were very important or critically associated with some function in our body, that being one of the examples. And then saying, oh, out of X, Y, and Z, Y turns out to be statistically significant. And so therefore we're able to predict some function at, I don't know, 90% accuracy. I would want to know more about what is the interaction between X and Y? What is the interaction between Y and Z? Was Y standalone important clinically? Or is there something going on in the dynamics between the features that is clinically meaningful for me? So it doesn't matter too much that I have a model that's 95% accurate versus 90% accurate, but if the 90% accurate model is able to touch upon these featured interactions and other clinical questions I have, I might find that a little bit more useful.	2-5-1	Useful - Useful for task
		13	And it is more or less like a sum of different individual effects. Everyone who applies such a model can see that this contribution comes from age, this contribution comes from hypothermia, this contribution comes from low blood pressure and so on and so on.	2-5-1	Useful - Useful for task

Section	Visualization tasks	ID	Text	# Need	Theme - Sub theme
	Assessing models at different cut off or thresholds	20	So if you say, this is what the physicians are doing now. And if they use this model at a particular cutoff, this is what would have happened. So right now they're treating 30% of the patients, right? And if they were to use this cutoff, they could reduce that to treating 20% of the patients. So they would treat fewer patients. And what would the trade-off be in terms of the patients who would benefit from treatment? So when they're treating 30% of the patients, only 10% of them benefit. But they could reduce to treating 20% of the patients and maybe 20% of them would benefit. You'd have to almost like show it in pictograms or something for the doctors to understand it.	1-6-3	Understandable - Evaluation
		11	But sometimes, in certain diseases or certain treatments, you know, already, right? I don't know, for example, we say, if the risk of having a new blood clot, for example, is higher than 5% over a year, then these patients usually benefit from staying on treatment to prevent such a blood clot. So, you know, some situations where you know the anchors, like you know, the thresholds above which or below which such an intervention becomes desirable, and others where it gets tricky.	2-1-1	Useful - Critical support
		7	Yeah, the net benefits as they call it. I think there is a direction towards that step to throwing the usefulness. I think it is the intention. But I do not necessarily agree that whether it is really useful, it is another debate. But at least I think the intention behind that is that it tries to see at least at what thresholds of the predicted value, the prediction models can be useful.	2-5-1	Useful - Useful for task

Section	Visualization tasks	ID	Text	# Need	Theme - Sub theme
Model performance	Reporting model performance using multiple metrics	18	So for example, when it is an statistics and data science teams, you are most likely to see both discrimination and calibration. But when it is mostly a physician team, you see a lot of sensitivity, specificity..	1-2-1	Understandable - Comprehension
		12	I know people like ROC curves and things like that. I think the average clinician, they find that those quite tough unless they're very clearly marked. So, yeah, whereas calibration, that the lines are far easier to understand than the text. So, as an example, I'd say there are some things that are very clearly visual, like a calibration line, and some things that need a good example, very clear non-technical example. And I'd say a discrimination statistic is in need of that.	1-2-1	Understandable - Comprehension
		3	But in medical literature, that is commonly accepted that you should have at least area under the curve as a performance measure. But if you say that you have not reported area under the curve, it is a pain in the neck.	1-3-3	Understandable - Analysis
		7	Maybe. The performance. But the performance I mean the calibration not necessarily the discrimination, because I think when I look at just the ROC curve doesn't, doesn't mean that much. It can mean that I think the value of the AUC itself is enough to my, in my view. Maybe I'm wrong.	1-3-3	Understandable - Analysis
		13	This is sometimes. Some call it C statistics and others call it the area under the ROC curve. I would prefer the last one. Even a normal doctor wouldn't know what an ROC curve is, but this is some kind of basic stuff which you need to know when you work with prediction. Okay.	1-3-3	Understandable - Analysis

Section	Visualization tasks	ID	Text	# Need	Theme - Sub theme
		20	Well, I mean, I think that, you know, having calibration plots are helpful, you know, and the AUC curve is helpful for people that know what it is, but it is not helpful for people who don't know what it is. And yeah, often you don't see anything about calibration, which is obviously a very important piece of it.	1-3-3	Understandable - Analysis
		14	So we think of the C statistic in terms of discrimination, or we think of Breyer scores, or we think of, what am I, I'm thinking of that graph. I'm thinking of the accuracy of the model, any indices that we use from that.	1-3-3	Understandable - Analysis
		19	The other aspect of clinical prediction models is that there must always be a calibration curves. And so I insist on them because otherwise people think they can get a probability and it is meaningful. And that may have a very good discrimination but poor calibration.	1-3-3	Understandable - Analysis
		1	I think they're called decision curves. Yeah. I think that's what they're called. And I started to look at those at one point because I remember a reviewer suggested that these were the solution.	1-3-3	Understandable - Analysis
		6	Yeah, I think that for kind of basic like performance summary certainly like an ROC curve and a calibration curve or something that, you know, showing that, like, of the, of the patients or whatever you say, have a 10% risk, what's the actual empirical probability or something like that.	1-3-3	Understandable - Analysis

Section	Visualization tasks	ID	Text	# Need	Theme - Sub theme
		7	And then a fair presentation of the performance as well, as I said before. I think that's that's important to see if the, if the, if the prediction, the predictive values agree with the observed value of the outcome. So, it is important.	1-3-3	Understandable - Analysis
		11	or if it is a, if it is a score, you know, the area under the curve helps sometime understanding how, you know, whether it is more like a sensitive or specific or, you know, things like that.	1-3-3	Understandable - Analysis
		18	Sometimes most often, I think they, for example, only report the discrimination, but not the calibration or the other way around. But fewer and fewer studies, you find both. And sometimes what I feel is like results reporting is sort of incomplete. And I think it also has to do a lot with the background of the authors	1-3-3	Understandable - Analysis
		1	What I feel like they haven't done is they haven't carefully shown the reader the trade-offs that that prediction model can have. So they haven't thought about, well, is it highly sensitive and not very specific? Or is it very specific and not very sensitive? And they've just bundled everything together. They've got one measure of accuracy, the C statistic, but they haven't thought about the trade-offs. And I think that's what I see as the major concern.	1-3-3	Understandable - Analysis
		2	I guess I'd like to see, if not replication, then I'd like to see some sensitivity analysis.	1-3-4	Understandable - Analysis

Section	Visualization tasks	ID	Text	# Need	Theme - Sub theme
		20	Like, yeah. I mean, I guess it is still like a lack of imagination. But I still think that the same, you know, the same metrics, you know, error, and the ROC curve, prior scores. And one of us is frozen. And calibration plots are really important. You know, particularly the calibration, because the calibration can be way off. Even if the discrimination is really good, the calibration can be terrible.	3-2-4	Trustworthy - Objective
		6	Yeah, it could be. Yeah, yeah, accuracy, MSC calibration, some, some, you know, some different metrics, you know,	3-2-4	Trustworthy - Objective
		6	But, you know, and you know, do they have the minimum, kind of like do they provide like the AUC or like some performance metric do they provide that. Do they actually give you enough that you can evaluate. If this is trustworthy.	3-2-4	Trustworthy - Objective
		20	Um, I don't know. I think it is just the things that they omit that make it less understandable. That is to say, if you don't include a calibration plot, people may not understand that the model won't be calibrated for their own data. So I could, I could show just the discrimination in a, in another dataset. You know, I could show that I validated it in an external dataset and that the discrimination is similar to what I saw in my original dataset. The discrimination might be the same, but the, but the calibration might be really different. And so, you know, unless I know what that is, I can't use the formula that they derived to come up with the probability of somebody having that outcome in my data. Okay.	3-3-2	Trustworthy - Validity

Section	Visualization tasks	ID	Text	# Need	Theme - Sub theme
		18	I mean, you could report discrimination and collaboration and then somehow find the perfect cutoff point and then report sensitivity and specificity and all those things that really matter for clinicians. But that's not that important from the risk prediction modeling point of view.	1-3-3	Understandable - Analysis
	Presenting more than just calibration plots	7	Second, when the figures that are reported are not the easiest to interpret. Like, I see some papers that have reviewed where they don't present the calibration curve, instead they just present the ROC. In terms of calibration, I was not able to see in what region of the predicted probabilities, there was a mismatch with the observed probability so that I was not able to see because the calibration curve was not shown.	1-3-3	Understandable - Analysis
		14	And then sometimes it turns out that you are more your predictions are more accurate in the lower range or more accurate in the high risk range and less accurate in the middle. And so that can be helpful in helping people understand how closely the model fits and how much how precise those estimates might be so yeah there are there are graphics that you typically see that can be that can be helpful but again it just, it depends on who the audience is and how much they want to read about this stuff and get into it.	1-3-3	Understandable - Analysis
		3	But that rigorous testing is based on the calibration and the very few studies in the literature actually assess the calibration very well. They just plot the calibration and they said that's done. But in fact, assessing those calibrations and the results and to understanding like how this varies and do you need some kind of recalibration of the model depending on	2-5-1	Useful - Useful for task

Section	Visualization tasks	ID	Text	# Need	Theme - Sub theme
			the that side or depending on the outcome in a different sides. .		
		18	Even if it is a bar chart or you know, a scatterplot between predicted and observed risk. That's, that's, that's it. I think if we could stick to those three parts in our table clearly reported the coefficients and predictors. Some few reports in your discrimination metric, even, even just a sentence, and something You can report one metric, but that doesn't really tell you whether there's underestimation or overestimation and whether, and where it is happening. You know, you can be underestimating the lower risk and overestimating at the upper risk. But then that's why these figures are important and I don't often see those.	2-5-1	Useful - Useful for task
		19	Yeah. Yeah. The, it depends what you're doing, but as I said, you must present, and it should be part of the main paper, not the supplement, the calibration curves. They also must interpret them. I've seen calibration curves, and then people say, oh, they're very good. Except when you look at it at the low end where it matters, perhaps, they're not very good. They're out by 50%, or something like that. But over the whole thing, they might look okay.	1-3-3	Understandable - Analysis
	Comparing model performances	19	The other, I was thinking other graphical, yeah, other graphical devices, things like the net benefit, which can take into account the weighting, different weighting between false negatives and false positives. I don't see that very often, but that can help explain where something might be useful more than, say, standard practice or other things. And that was the other thing that I think that is often missing with some	1-6-3	Understandable - Evaluation

Section	Visualization tasks	ID	Text	# Need	Theme - Sub theme
			types of prediction models is they're not comparing it against standard practice. And so you don't know if it is improving.		
		19	I tried to explain a new metric that was being used called the net reclassification improvement metric, NRI. And it was being used wrongly in the field and it was a kidney field I was working with nephrologists. But I also showed another graphical way of explaining how one model can improve. And for example, it might improve the predictions for those who had the outcome, but it may make no difference for those who didn't have the outcome in terms of the probabilities that came out of the model. So, the first model, maybe somebody had a probability of 0.5 of dying within six months and they did die. And the second model might be 0.6, so that's an improvement. But somebody who didn't die had a probability of say 0.2, but in the new model, they also had a probability of 0.2, so that didn't improve. And so it can show those separately improvements. So, I've seen that	1-6-3	Understandable - Evaluation
		6	Or that's, you know, not exactly but it is possible that you could have two models that predict equally well that have three only a small number of covariates that are completely felt like the over, they don't overlap.	1-6-3	Understandable - Evaluation
		19	Would it be possible if they're doing, say, a box and whiskers type of plots to compare a couple of things? I say also put the underlying points in the graph. I've actually made some of my own types of graphs to try to show improvements and predictions between two models.	1-6-3	Understandable - Evaluation

Section	Visualization tasks	ID	Text	# Need	Theme - Sub theme
		18	If you're reading a risk prediction model paper and you are thinking of using that in your patients or in your practice or for something what pieces of information do you look for to make your decision of using model A or model B? And once you answer that for yourself we should try to always report that. And I think the best place for that is probably the methods section where we can clearly describe the population and clearly describe the variables as I told you at the beginning many, many risk score papers I just can't figure out what the predictors were.	3-2-4	Trustworthy - Objective
		6	Then, assuming that they've provided all of that then I think you had start going into looking at the performance, and also what the baseline. Now, this is something I don't think it is done very often but I'm curious, would be curious is if there's an area where there's kind of a baseline prediction model or an already used tool, like how does this model compared to the baseline.	3-2-4	Trustworthy - Objective
		6	Yeah, but it is ideally, you should be showing not just that you do a good job you should be showing that we do like a better job, or we do a comparable job but our model is simpler, you know, which I don't think is the norm. From my experience reading literature.	3-2-4	Trustworthy - Objective
Model presentation	Presenting simplified model presentation	6	Is that sufficient to be understandable, or does it have to be a really simple rule, like, you know, a simple decision list or some sort of scoring system where it is really easy, or it is understandable both what goes into it and how the, how the model is operating.	1-4-1	Understandable - Application
		6	But they do also kind of have the issue to me that if that's not the actual model that you're using. You	1-4-1	Understandable - Application

Section	Visualization tasks	ID	Text	# Need	Theme - Sub theme
			know, if you're just creating this nomogram to kind of explain or make it easier to see what the model is doing.		
		11	Well, again, so maybe that Fagan's diagram or showing how, for example, I don't know if a certain disease requires treatment above, like putting that in perspective with the next steps again, right? Whether it is a diagnostic test and then it could be a Fagan's nomogram of how such a pre-test will be influenced by the result of the diagnostic test and what will be the post-test probability of disease or some discussion of risk benefit of treatment and above which probability of disease the chances is that patient will get more benefits at harm from being treated for that disease.	1-4-2	Understandable - Application
		18	I think it has to do with what we were discussing initially, like just report everything and, you know, maybe use a nomogram or it is not ideal, but maybe translator or research scoring to a point based score, something where people could actually do. I mean, if you just report a set of coefficients, I think you can't expect people to actually do the math and compute themselves the absolute risk. That's not going to happen.	2-2-1	Useful - Easy to use
		18	If that is your aim, if that is your target population, I would use a risk chart or a nomogram or translate your model into a point-based model. If you have this at one, you have that at two, and so on. It is not ideal from the statistical point of view. At least it is going to be usable or most likely to be usable in those contexts.	2-2-1	Useful - Easy to use

Section	Visualization tasks	ID	Text	# Need	Theme - Sub theme
		18	Yeah. I don't know if I will use the word appealing, but I will use the word useful. Useful. Or usable. Otherwise, we can have a very nice paper and report everything transparently, but again, you're requiring the other person to compute them. Take your cell phone and manually compute the risk. That's just very, very difficult. You can do something else like a risk chart. You can have it print or some tool that actually makes my life easier as an end user.	2-2-1	Useful - Easy to use
		19	But I do an icon array where I've got little pictures of people and however many, depending on the probability. Now, some clinicians go, oh, I can see that whereas, 17.4% doesn't quite work for them, but they can see it and patients can see it. So there can be a discussion.	2-2-1	Useful - Easy to use
		6	Okay, yeah, I want to say that I had come across a paper that I at least was sort of using as a reference that I like some of what they what they did. And I know that sometimes what I've seen a few times what authors will do is kind of back fit. What is it called a nomogram, something like that, you know what I'm talking about. nomogram. Where they kind of back fit one of those to their model, even. So they have a complicated model and then they have simpler scoring.	1-4-1	Understandable - Application
	Optimizing model presentation	6	You can't show what the entire space of predictions looks like, apart from letting somebody play around with the calculator.	1-4-1	Understandable - Application

Section	Visualization tasks	ID	Text	# Need	Theme - Sub theme
	through interactive visualization that engage the users	6	Yeah, so that's, that's something, you know, where users could play around. I'm also thinking of if you're, you know, writing a paper and you're stuck with the non-interactive format to give a contract that contrived example, you know, if we have some risk prediction model where the only things are age and number of comorbidities, you know, you could either fix the number of comorbidities, and then very age, but that's not. That's a little misleading, in the sense that that's not what very old or very young people really look like. And so, you can vary age, and it each age have whatever the, what the number of comorbidities be like the mean number at that age. And that might be more.	1-4-1	Understandable - Application
		7	I think. I think it should be close to for the presentation of the model I think should be close to what the user would be provided with. Like, I like the idea of, and some of the papers already showed that. Like, web, web based calculator, for example, online calculator. It is nice in a way that one can see the different items. One can see how to fill in the different items, and then one can see that is written something. That's. So that's one thing.	1-4-1	Understandable - Application
		9	So they know what the inputs are basically. I don't know, like I'm making this up, like say it is, you know, like, so sometimes they use calculators on their own and when they put in stuff into the calculator, I don't know, patient's blood pressure, patient's this measure, patient's that measure, they know what's going in, right? So they're able to trust it, understand	1-4-1	Understandable - Application

Section	Visualization tasks	ID	Text	# Need	Theme - Sub theme
			whatever calculator they're using and they understand they can trust it because they know how it works.		
		10	And then the second thing is I guess afterwards it is about informatics using a web app can help sometimes to have a better understanding. I guess that'll be that.	1-4-1	Understandable - Application
		14	And so, we're going to. And then we created. We took that information and then we created a website with a with a clinical prediction calculator in it. And so, we incorporated all this information, created a website, and then our next thing to do would be to go out and validate that at the impact on and we're doing a pilot study right now at Iowa, trying to look at the impact of getting that information on what impact does it have on the patients. Are they more compliant with treatment. If they saw these risk estimates, and then what impact did it have on physician prescribing practices, too.	1-4-1	Understandable - Application
		14	Oh, okay. Like calculator or handout that might have, you know, those pie charts or the line graphs or whatever you think is going to be most understandable to have those. And then the patient has something in their hand. Yes. So, and I can send you examples of some stuff that we've developed and we're testing right now.	1-4-1	Understandable - Application

Section	Visualization tasks	ID	Text	# Need	Theme - Sub theme
		14	Then the second step is that the clinician has to have enough information about what the implications of all of this are. So they have to understand that. So you have to provide them with that information. And then I think it would be great if you gave them tools to make this easy for them to use. So creating online calculators or a website or something that they can easily access in clinic.	1-4-1	Understandable - Application
		14	But then you also should be also helping them to be able to then use that information in a conversation with a patient appropriately. Okay, so if you can then go the next step and create a handout, create a website or create something that they can use in the clinical interaction, that way you're also kind of protecting your own work, right? Because people misinterpret modeling, people misinterpret research all the time, right? And so if you say, okay, here's my model, this is what it means and use this handout, then you have also then I think you're keeping more control of your work and you are going to decrease the risk that is going to be misused or misunderstood by both the clinicians and by patients.	1-4-1	Understandable - Application
		18	Yes, it should be understandable, but also complete. Because sometimes you can find like, I don't know, the coefficients, but not the baseline risk or the intercept. So basically, unless they put together a software or an Excel sheet or something like that, then you basically cannot use that risk score. Basically, it has a set of coefficients, but not the baseline risk or the intercept. So you cannot replicate it.	1-4-1	Understandable - Application

Section	Visualization tasks	ID	Text	# Need	Theme - Sub theme
		6	You know, if you have like a app or something and it is asking for the data that they think is clinically important, and it is spitting out predictions that make sense to them if they played around with it, you know, like if I am a clinician and I put somebody with more comorbidities at older age into the risk prediction model, does the risk go up or down, you know, like, in my very limited experience, you know, when you give somebody sort of a prediction modeling tool, you know, what they like, what I've seen people do is they play around with it and see if it produces results that are kind of in line with their thinking. And so, or at least make sense to them, you know.	1-4-1	Understandable - Application
		10	All right. So, of course, when you look at, let's say my work is on asthma attacks, there are different risk factors for asthma attacks. There's asthma attack history, which is the most important one, intense severity of disease and then biomarkers. OK, so that's it. That's an example. You could make an equation based by points and everything and then use it as a MedCalc.	1-4-1	Understandable - Application
		10	Is it usable because there's a web app? And we made a web app with our little score. So we think the web app part is a very important part. MedCalc is another very important part. So that's how I see it.	2-2-1	Useful - Easy to use
		13	What we publish is a kind of rather simple model. This is one of our models for example. There is a plus and minus calculation. With a calculator or with a computer, you know how the prediction is done. This is open to everyone. This is clear how our prediction	2-2-1	Useful - Easy to use

Section	Visualization tasks	ID	Text	# Need	Theme - Sub theme
			model works and how the different observations or findings were included in the final result.		
		18	I think to improve the usability of the model, we have to move away from the paper itself. The paper has to be really clear and show all the information. Make a tool, something, whatever that is, you can actually do for people to actually use it. Otherwise, it is just going to be difficult for your model, for any model, to hit the practitioner's desk or anyone at the desk.	2-2-1	Useful - Easy to use
		18	No, I think we have to think when doing a research either diagnostic or prognostic. Who do we want it to be used by? Who is the target audience? Who will be the end users? Based on that, we can decide on the best tool. If you are developing a model and you expect it to be useful or usable for low and middle income communities or primary care settings where they do not have computers, for example, I would not do a Shiny app or any other app.	2-2-1	Useful - Easy to use
		18	So, you know, very simple thing, maybe just upload as a supplementary material your model as an Excel sheet. That's rather simple. If you have the knowledge or the skills or the time, maybe just put sort of a shiny app if you work with R or a plotly app if you work with Python or whatever, or any other similar tools.	2-2-1	Useful - Easy to use
		19	So the other aspect of usefulness is, and I've seen things like Shiny apps or things on the web where people can put in the information and get a probability or a graph. If you look at that wayfind.health that I just sent you, that's got a demonstration there of a clinical pathway for chest pain and behind it is a little model, simple model of logistic regression.	2-2-1	Useful - Easy to use

Section	Visualization tasks	ID	Text	# Need	Theme - Sub theme
		20	Those kinds of models, especially if they've been around for a long time and they've been validated in different study sets. You have something like the Chad's VASC, which is like just add up these points. Or you have the pooled cohort equation for coronary artery disease. That one seems to be pretty popular. That one sort of moved away from points into a calculator that gives you a risk. That's kind of where we should be. You select the risk factors and then it tells you what the probability is that something's going to happen. But physicians aren't good at working with probability.	2-2-1	Useful - Easy to use
Multi section	Incorporating confidence intervals	19	In terms of things like ROC curves or any curves, certain things I insist on is that they have some measure, usually a confidence interval presented on any graph.	3-2-1	Trustworthy - Objective
		13	Then it would suggest another figure like where the odds ratios are plotted with confidence intervals and maybe in a sequence that the most important is on top and so on and so on. So this is more or less to say what importance do certain predictors have. In that case, it is not the primary aim to have an exact probability at the end, but to say what is the adjusted importance of some predictor. Then I would choose some different figure.	3-2-1	Trustworthy - Objective
		19	But, the trustworthiness. How can it be produced? I think what's really important, and in fact is more important than the point estimates, is the, what I call measurement error, which is usually a confidence interval of some kind. And that's also shown, maybe shown with a graph sometimes. But, it is that, when I	3-2-1	Trustworthy - Objective

Section	Visualization tasks	ID	Text	# Need	Theme - Sub theme
			see a result, and sometimes I see, I often see this in an abstract, where the results are presented only the point estimates, and not the confidence interval. The confidence interval is more important than the point estimate. Because that's just the maximum likelihood of something, but the rest of it is telling you, well, what could it be?		
		19	So sometimes what I see is very simplistic. They give an AUC, and that's about it. Or, and they, there's no calibration. And sometimes they interpreting it based on the point estimate. So just the AUC and ignoring totally the very wide confidence intervals they have. And so they're not looking at what is really plausible. So that means the interpretation, you know,	3-2-1	Trustworthy - Objective

VITA

Ivan Rahmatullah, MPH, MD, Ph.D.

PhD Graduate in Biomedical and Health Informatics, University of Washington, Seattle

Bio:

Ivan Rahmatullah is currently pursuing his Ph.D. in Biomedical and Health Informatics at the University of Washington, Seattle. With a Master of Public Health from Emory University and a background as a trained physician and lecturer in Public Health at Universitas Airlangga in Indonesia, Ivan's expertise encompasses preventive medicine, health promotion, clinical informatics, and human-centered design.

Ivan has made significant contributions to the field of biomedical informatics during his pursuit of a PhD. In his role as a Graduate Research Assistant with the Digital Initiatives Group at the International Training and Education Center for Health (I-TECH) at the University of Washington, he has been instrumental in developing and promoting digital tools for clinical data capture, storage, analysis, and visualization, while incorporating human-centered design principles. His work with the Seattle Flu Study, analyzing flu and COVID-19 data, further highlights his adeptness at managing complex datasets. Ivan's academic contributions, including primary data studies and systematic reviews, have been published in various peer-reviewed journals.

At the heart of Ivan’s research lies his dedication to presenting data and information understandable, useful, and trustworthy. He is committed to improving decision-making processes for diverse populations by ensuring that complex information is accessible and actionable.