Development of an LLM Framework for Clinical Hypothesis Testing using Multimodal Data

Nayoon Gim

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington
2026

Reading Committee:
Ruikang K. Wang, Chair
Aaron Y. Lee, Chair
Russell N. Van Gelder

Program Authorized to Offer Degree:
Bioengineering

University of Washington

Abstract

Development of an LLM Framework for Clinical Hypothesis Testing using Multimodal Data

Nayoon Gim

Chairs of the Supervisory Committee:

Ruikang K. Wang

Aaron Y. Lee

Department of Bioengineering

Electronic Health Records (EHRs) contain rapidly expanding volumes of structured clinical data with the potential to accelerate evidence generation. However, translating clinical hypotheses into reproducible research remains a slow and resource-intensive process requiring manual cohort definition, data harmonization, and statistical coding. These labor-intensive steps limit scalability and transparency, contributing to challenges in reproducibility and auditability. This thesis investigates how large language model (LLM)-assisted workflows, combined with data standardization and privacy-preserving system design, can transform clinical research into a more scalable and transparent process. Chapter 1 introduces the motivation for LLM-assisted scientific workflows, reviews standards and interoperability challenges in health data, and defines the scope of the thesis.

Chapter 2 establishes the data foundations that enable automated LLM-assisted clinical research by addressing two complementary requirements: data standardization and secure LLM interaction with health records. Clinical datasets are often fragmented and inconsistently structured, leading to dataset-specific analytic code that limits scalable automation. We address this by standardizing retinal imaging

data in the AI-READI (Artificial Intelligence Ready and Exploratory Atlas for Diabetes Insights) cohort using structured DICOM (Digital Imaging and Communications in Medicine) representations. Building on this standardized structure, we develop a metadata preparation workflow that enables LLM-assisted analysis without exposing patient-level data. By aggregating schema information and natural language representations of data elements, this workflow provides the contextual information required for LLMs to generate executable analytical code without accessing any patient-level data. These approaches are demonstrated using two datasets: AI-READI and NHANES (National Health and Nutrition Examination Survey).

To understand what aspects of clinical research can be effectively automated, it is first necessary to analyze existing manual workflows. Chapter 3 begins with a case study investigating the relationship between post-intraocular pressure elevation and the development of primary open-angle glaucoma using the IRIS Registry (Intelligent Research in Sight). This study was carried out using standard manual research workflows and serves as a representative example of real-world clinical research practice. Section 3.2 then builds on this work by shifting the focus from clinical outcomes to process analysis. Using the completed study from 3.1 as a reference, we examine the underlying research workflow to identify repetition inherent in manual pipelines, scalability bottlenecks, and recurring analytical steps. This secondary analysis highlights concrete opportunities for automation and directly motivates the development of an LLM-assisted framework to streamline hypothesis testing.

Chapter 4 introduces LATCH (Large Language Model-Assisted Testing of Clinical Hypotheses), a framework that automates the translation of natural language research questions into executable statistical analyses. LATCH combines an LLM-driven semantic component that maps hypotheses to explicit cohort definitions and data extraction logic with a deterministic statistical engine that ensures reproducibility and auditability. We describe the system architecture and validate the framework by reproducing a set of published studies on diabetes using the NHANES data, demonstrating that LATCH can generate end-to-

end analytical pipelines from natural language prompts without manual coding. We further characterize the system's operational limits through targeted stress testing and behavior under edge-case conditions.

Finally, Chapter 5 illustrates the application of LATCH in advancing biomedical knowledge. Beyond reproduction, LATCH enables extended analyses of existing studies, including cross-dataset generalizability testing between NHANES and AI-READI, temporal consistency evaluation, stratified analyses, and more granular exploration of prior findings. LATCH is also used to conduct exploratory, hypothesis-generating analyses of previously unexplored questions, including the identification of a nationwide vision-related trend in the diabetes population and associations between disease severity and retinal biomarkers using the AI-READI cohort.

This thesis presents a framework that combines data standardization, privacy-aware infrastructure, and LLM-assisted analytics to improve the efficiency and reproducibility of clinical research. The work demonstrates that carefully designed AI-assisted systems can accelerate hypothesis testing, reduce repetitive manual effort, and support transparent real-world evidence generation while preserving human expert verification.

Table of Contents

Acknowledgements

I extend my deepest appreciation to my advisor, Dr. Aaron Y. Lee, for his constant support and mentorship throughout my PhD journey. I am truly grateful that he accepted me as his PhD student years ago after I started volunteering as a medical student back in 2022. His exemplary dedication to initiative, productivity, and perseverance has been a defining influence on my development as a researcher. I am especially thankful for his steadfast support and responsiveness, which have played a pivotal role in shaping my career trajectory.

I am sincerely thankful to my co-advisor and co-chair, Dr. Ruikang Wang, for making it possible for me to pursue my PhD in the Bioengineering Department at the University of Washington. I am grateful for the opportunity to work as a teaching assistant for the first time and for his continued support and guidance throughout my time in Bioengineering, which made me feel encouraged and well supported.

My sincere thanks go to my supervisory committee members, Dr. Van Gelder, Dr. Rubinstein, and Dr. Cherry (Graduate School Representative), for their thoughtful mentorship and for dedicating their time to reviewing my work. I am especially grateful to Dr. Van Gelder for his advice on prioritizing research projects and directing my focus, which ultimately helped shape my PhD research direction. I am especially grateful to Dr. Rubinstein for serving as an inspiring physician-scientist role model as a fellow graduate of the University of Washington's Bioengineering Department and for his continued support of our program and its MSTP students. I also sincerely appreciate Dr. Cherry for his willingness to serve as GSR despite his schedule and for his encouragement since my General Exam. Their constructive feedback played a vital role in completing this research, and I am thankful for their commitment to attending my defense.

I am privileged to have collaborated with exceptional mentors. I warmly thank Dr. Cecilia Lee of the Lee Lab, whose intellectual curiosity about clinical findings, expertise in biostatistics, and supportive leadership greatly shaped my development. I sincerely appreciate her for including me in collaborative research and publications that enriched my academic experience.

Special recognition goes to my colleagues in the Computational Ophthalmology Lab and to all collaborators whose support and friendship enriched my PhD experience. I am deeply grateful to Dr. Yulie Jiang, Dr. Yuka Kihara, Dr. Yelena Bagdasarova, Dr. Yue Wu, Dr. Marian Blazes, Jamie Shaffer, and Dr. Julia Owen, along with my many other colleagues, for their encouragement and assistance throughout my academic career.

I would also like to recognize Sanjay Soundarajan and Bhavesh Patel from the California Medical Innovation Institute for their meaningful professional and personal contributions to the AI-READI project. I wish them continued success in their future pursuits.

Above all, I offer my heartfelt thanks to my parents and my husband, a fellow PhD student, for their unwavering encouragement and support throughout my PhD studies. Their belief in me made this achievement possible.

List of Figures

List of Tables

Chapter 1. Introduction

1.1 Significance and Research Motivation

Despite the growth of data within Electronic Health Records (EHRs), the translation of clinical insights into evidence remains resource-intensive, requiring manual coding for cohort definition, data harmonization, and statistical analysis[1–3]. These labor-intensive steps prolong the research cycle and contribute to the reproducibility crisis, as the exact analytic process is rarely transparent or auditable end-to-end [4–7].

Prior attempts to streamline this process have revealed important limitations. Rule-based frameworks, while structured and explainable, often rely on rigid, dataset-specific representations that make it difficult to express nuanced clinical hypotheses, such as defining patient cohorts using combinations of diagnoses and laboratory criteria, and require substantial manual coding effort [8,9]. There remains a need for a framework that reduces the manual coding burden while preserving methodological transparency, enabling efficient hypothesis testing by a broad range of researchers.

While healthcare applications of large language models (LLMs) have focused on tasks such as clinical literature summarization[10,11], patient engagement[12,13], or clinical decision support[14,15], the use of LLMs to support research analytics, translating natural language hypotheses into executable code, remains relatively underexplored. This thesis introduces LATCH (Large Language Model-Assisted Testing of Clinical Hypotheses), a privacy-aware and auditable framework that serves as an interface between well-specified natural language research questions and deterministic statistical analyses over standardized data.

Prior work has shown that AI approaches operate as opaque systems, limiting verifiability and explainability[16–19]. In contrast, LATCH does not use AI to produce predictions or conclusions. Instead, the LLM is used solely to translate natural language hypotheses into explicit, executable analysis steps, shifting AI output from end results to an analytical process, all of which are fully auditable. Although human review remains essential to ensure methodological rigor, as with any LLM-utilizing application, the analytic report provided by LATCH enables practical human-in-the-loop verification by providing a detailed step-by-step record.

Within the LATCH framework, the researcher remains in control of the scientific inquiry, providing both the hypothesis and the desired analytical method. LATCH translates the query into a series of database operations and statistical analysis by mapping concepts to the database schema and generating the necessary code. Importantly, LATCH is a privacy-aware framework as it only exposes the database schema (e.g., table and column names) to the LLM, while strictly isolating all patient-level data from any LLM-involved steps, addressing privacy concerns regarding LLM use[20–22]. LATCH is designed to augment domain expertise rather than replace it, enabling researchers to focus on scientific reasoning, hypothesis formulation, interpretation, and verification rather than manual implementation and debugging.

This thesis investigates how this framework can be utilized for reproducible real-world evidence generation, reducing analytical bottlenecks and improving reliability of AI-assisted biomedical discovery

while preserving human oversight. Portions of this chapter are adapted from the author's preprint[22] and publication[23].

1.2 Standards and Interoperability in Health Data

DICOM (Digital Imaging and Communications in Medicine)

DICOM is an international standard used to manage, transmit, and save medical imaging information in healthcare systems[23]. It was developed to facilitate the integration and interoperability of imaging systems, allowing communication between devices such as magnetic resonance imaging (MRI) and medical imaging management software including Picture Archiving and Communication Systems (PACS).

A DICOM file combines image data (such as pixel data) with extensive metadata, which includes information about the patient, the study, the modality, and imaging parameters like resolution or contrast settings. This structure enables DICOM files to provide not just images but also detailed contextual information needed for appropriate interpretation and application of medical images.

DICOM also specifies a communication protocol for transmitting imaging data over networks, ensuring that imaging devices and systems from different manufacturers can communicate. The protocol supports querying devices, retrieving studies, and sending images between systems using standard messages. This interoperability is a cornerstone of DICOM's widespread use in the medical imaging field.

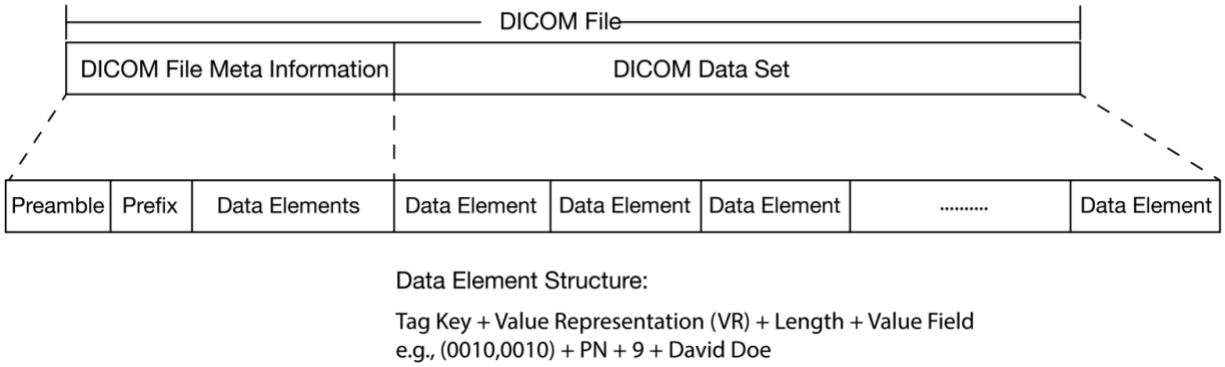


Figure 1.2.1. DICOM file structure. Reproduced from ref. 23 under a CC BY 4.0 license.

As shown in Figure 1.2.1, a typical DICOM file includes a structured File Meta Information Header, which consists of a File Preamble (a 128-byte field reserved for compatibility with non-DICOM systems), a DICOM Prefix (a 4-byte signature identifying the file as DICOM), and the File Meta Elements, which are Data Elements in the DICOM File Meta Information. The File Meta Elements include critical metadata, such as the Service-Object Pair (SOP) class unique identifier (UID), transfer syntax, and implementation version. DICOM data is stored as a collection of Data Elements, each with a specific format: a Tag (unique identifier indicating what the data describes), a Value Representation (specifying

the data type such as binary or string), a Value Length, and the actual Value (such as pixel data or patient name). The image data itself, known as Pixel Data, is also stored as a Data Element. In addition to handling simple images, DICOM can store multi-frame data, such as a series of optical coherence tomography slices, in one file.

Unlike common image formats such as PNG or JPEG, DICOM is specifically tailored for medical imaging. It offers comprehensive metadata that includes patient demographics, study information, and technical parameters. DICOM also supports multi-frame imaging, imaging compression options, and advanced imaging-related features such as multi-dimensional datasets and annotations.

DICOM represents imaging data as Information Object Definitions (IODs), which are specific to imaging modalities. For example, an MRI scan has a different IOD than a CT scan. The DICOM standards document, PS3.3[24], contains a variety of IOD modules with definitions of the information objectives, description, and the required modules, which specify a list of tags needed, and their tag value requirements.

Additionally, because DICOM files contain extensive metadata including personal identifiers, necessary attention must be paid to anonymization. Unlike standard image formats, which may have limited or no embedded metadata, DICOM files can contain multiple tags with potentially patient identifying information. Even when obvious fields like Patient's Name (0010, 0010) are removed, other tags such as Patient's Birthdate (0010,0030), Ethnic Group (0010, 2160), Sex (0010, 0040), Age (0010, 1010), Occupation (0010, 2180), Address (0010, 1040), and Phone Numbers (0010, 2154) may remain. For these reasons, anonymizing DICOM data requires not only stripping standard identifiers but also a systematic review of tags, especially when datasets are used for research, AI development, or public sharing. Portions of this chapter are adapted from previously published work[23] by the author.

Observational Medical Outcomes Partnership (OMOP)

The OMOP Common Data Model is an international standard for representing structured clinical and observational health data in a consistent, research-friendly format. Developed by the Observational Health Data Sciences and Informatics (OHDSI) community[25], OMOP was designed to enable large-scale observational research by harmonizing heterogeneous healthcare datasets into a shared schema using standardized vocabularies. By converting institution-specific data into this standardized structure, analyses can be run consistently across different datasets without repeating extensive data standardization for each study.

In the OMOP model, clinical information is organized into standardized tables representing core domains such as observation_period, death, visit_occurrence, condition_occurrence, observation, measurement. Each table follows a predefined schema with consistent field names and relationships, enabling interoperable querying and analysis. A key feature of OMOP is its use of standardized vocabularies, which map local coding systems to shared clinical concepts.

In addition to core clinical domains, OMOP includes tables that capture administrative and contextual information. For example, location, care_site, and provider describe aspects of the healthcare delivery system, while cost and payer_plan_period support standardized health economics analyses. These

domains are relationally linked to the core clinical tables through keys, allowing integrated analyses across clinical, organizational, and financial dimensions of care.

Transforming source data into OMOP, however, requires substantial curation and oversight. Source systems differ in coding practices, structure, and data completeness, and mapping to standardized vocabularies can introduce uncertainty. In addition, many OMOP datasets contain sensitive patient information and must be managed under security requirements. When implemented carefully, OMOP provides a strong foundation for interoperable clinical research and supports reproducible, large-scale analytic workflows.

1.3 Emerging LLM-based Agentic Workflows for Assisting Scientific Discovery

LLMs are a class of deep neural networks with large parameter counts (millions to billions to trillions) trained on large-scale textual corpora to learn statistical representations of language. Although the term lacks a universally standardized definition, it generally refers to high-capacity models trained via self-supervised objectives at scale[26]. Architecturally, most modern LLMs are based on transformer networks, which use self-attention mechanisms to model long-range dependencies and contextual relationships within sequences[27]. Through large-scale pretraining, LLMs acquire the ability to perform a wide range of tasks, including summarization, reasoning, and generative writing, often without task-specific retraining[28]. More recently, LLMs have evolved beyond purely linguistic tools to serve as reasoning engines capable of planning, tool use, and multimodal integration. These capabilities position them as promising computational partners in scientific discovery and biomedical research.

LLMs as Workflow Coordinators in Research Pipelines

Across recent work, the flexibility of LLMs to take and produce natural language has shifted how they are positioned in research systems. Unlike traditional supervised machine learning models that are trained to predict a specific label or numeric outcome, LLMs can take in open-ended instructions and generate structured text or explanations as output with reasoning capabilities. Because of this, they are increasingly framed not as standalone answer engines, but as orchestrators that coordinate multi-step research workflows, connecting problem formulation, information retrieval, analysis, and result generation within interactive loops.

For example, one of the earlier notable works is chemistry automation[29], where an LLM does not just describe experiments but coordinates the whole workflow by searching for reaction information, consulting instrument documentation, running calculations, and generating executable protocols for automated labs. In data-centric science, DataVoyager[30] advances a similar vision by presenting a conceptual framework in which an LLM agent could interpret a dataset, propose hypotheses, generate and execute analysis code, and evaluate results within a structured loop. However, the work is primarily a position paper that articulates this research direction and demonstrates feasibility in limited examples, rather than reporting large-scale autonomous deployment or sustained real-world execution. Extending this idea further, the autonomous research agent framework[31] describes multi-agent systems that decompose the research process into sequential, human-verifiable stages from data analysis to manuscript drafting, positioning the LLM's primary contribution as coordinating tools, intermediate outputs, and validation steps rather than producing a single final answer. Across these efforts, LLM-based research

systems tend either to propose agentic workflows conceptually or to showcase limited case studies, suggesting that the central innovation lies in treating the model as a controller of structured processes rather than as an isolated generator of text.

LLMs for Specific Research Task

At the same time, there is a growing body of work that applies LLMs to specific, well-defined research tasks rather than full end-to-end orchestration. Many of these applications focus on literature-centered workflows, since scientific publications are already text-based and naturally aligned with the strengths of language models. In systematic reviews and meta-analyses, LLMs have been used to assist with publication search, title and abstract screening, and data extraction[32]. Studies show that LLM-supported screening can significantly reduce reviewer workload while maintaining recall levels comparable to manual curation[33]. Toolkits such as LitLLM[34] and related systems demonstrate how LLMs can structure and accelerate literature review pipelines, while other work explores automated support for study selection and information extraction.

Beyond screening and extraction, LLMs have also been explored as research topic generators[35]. Some systems aim to synthesize patterns across large corpora of scientific papers to propose novel research hypotheses[36] or identify underexplored connections between concepts. These approaches are particularly relevant in translational medicine, where generating cross-domain insights between molecular mechanisms and clinical outcomes often requires integrating evidence across diverse literature sources. Overall, many current LLM applications in research operate as task-specific assistants, with literature search and synthesis being the most common and natural starting point due to the textual nature of scientific knowledge.

1.4 Scope and Contribution of the Thesis

This thesis investigates the design, evaluation, and applications of LLM-assisted workflows for automating hypothesis-driven clinical research. The primary focus is on translating natural language research questions into reproducible analytical pipelines operating on structured health data. The work emphasizes system architecture, data standardization, and workflow reproducibility rather than predictive modeling or clinical decision-making. LATCH is designed as an infrastructure for executing verifiable analyses, not as a system for generating new hypotheses or novel analytical methods. Instead, it operationalizes user-specified research questions by translating natural language hypotheses and selected statistical approaches into executable analytical pipelines using established statistical procedures implemented within the LATCH framework.

The scope of this thesis is restricted to structured clinical and observational datasets, including tabular health records and standardized medical imaging metadata. Most analyses focus on cross-sectional or relatively simple longitudinal datasets, reflecting a proof-of-concept setting for automated research workflows. Unstructured modalities such as free-text clinical notes, raw imaging interpretation, and real-time clinical deployment are outside the scope of this work. The framework is evaluated in research environments using publicly available datasets.

This work contributes to three interconnected areas: (1) data standardization and metadata preparation for enabling automated analytics, (2) the development and validation of the LATCH framework for translating natural language hypotheses into executable analyses, and (3) the application of LATCH to reproduce prior studies, extend existing analyses, and support hypothesis-generating clinical research. These contributions are organized across the following chapters.

Chapter 2 presents methods for standardizing imaging data and preparing metadata structures that support privacy-aware LLM interaction. Chapter 3 analyzes manual clinical research workflows to identify bottlenecks and recurring patterns that motivate automation. Chapter 4 introduces the LATCH architecture and evaluates its robustness and operational limits. Chapter 5 demonstrates the application of LATCH in exploratory, hypothesis-generating analyses.

Rather than seeking to automate all aspects of scientific hypothesis testing, this thesis presents a system focused on the interface between natural language study specification and executable analytical workflows. In this framing, LATCH is not positioned as a system that autonomously performs clinical research, but as a privacy-aware and auditable interface that translates well-specified research questions into deterministic analyses over standardized data. Its contribution lies in reducing implementation burden and improving reproducibility while preserving human control over initial study design, methodological judgment, and interpretation.

This thesis makes several contributions to AI-assisted clinical research infrastructure. First, it establishes a proof-of-concept framework for scalable and transparent clinical data analysis, demonstrating how contemporary language models can be integrated to accelerate hypothesis-driven research workflows while improving reproducibility and auditability. Second, it introduces LATCH as a model-agnostic system architecture in which large language models function as a semantic interface to deterministic statistical analysis, making the framework adaptable across different present and future models. Third, it develops a privacy-aware approach to LLM-assisted analytics through metadata abstraction and standardized data representations, addressing a core challenge in the responsible use of clinical data. In addition, this thesis demonstrates the multimodal potential of such a framework by showing how imaging-derived features can be incorporated as tabular data within a unified analytic workflow. Collectively, these contributions show that the value of this work lies in accelerating analytic implementation of clinical research, improving transparency, and augmenting human domain expertise within governed research environments, rather than in replacing expert judgment or focus on completely autonomous clinical research.

Chapter 2. Data Foundations and Governance for LLM-assisted Clinical Research

This chapter establishes the data foundations and governance principles necessary for automated LLM-assisted clinical research. Scalable research automation requires clinical data to be represented in standardized, machine-interpretable formats and accessed through mechanisms that prevent exposure of patient-level information to external LLM systems. This chapter addresses these challenges by presenting approaches to imaging data standardization and metadata organization that support reliable computational analysis while enabling secure interaction with LLM-based tools. Together, these methods provide the structural framework required for transparent, reproducible, and privacy-aware automated research

pipelines presented in Chapter 4. Portions of this chapter are adapted from the author's preprint[22] and publications[23] by the author.

2.1 Retinal Imaging DICOM Standardization Framework

2.1.1 Introduction

The demand for interoperability and the exchange of imaging data has long existed, particularly in fields where imaging plays an important role in advancing medical practice [37–42], such as ophthalmology, radiology, pathology, and dermatology. In ophthalmology, imaging modalities such as color fundus photography (CFP), infrared imaging (IR), fundus autofluorescence (FAF), optical coherence tomography (OCT), and OCT angiography (OCTA) support diagnosis, monitoring, and disease characterization. More recently, the rise of artificial intelligence (AI)[38–45] has intensified the demand for large, harmonized imaging datasets, because model training and validation often require combining data across devices, institutions, and protocols.

Despite widespread recognition of DICOM as the international standard for medical imaging, adoption in ophthalmology has remained limited. The American Academy of Ophthalmology (AAO) and the National Eye Institute (NEI) have both emphasized the need for standardization [46,47]. Adherence to DICOM standards would not only enhance data quality but would also enable large-scale collaborative research, leading to better clinical outcomes. The Collaborative Community on Ophthalmic Imaging (CCOI), composed of clinicians, patients, researchers, medical devices manufacturers, not-for-profit organizations, and government entities including the U.S. Food and Drug Administration (FDA) and the NEI, was created with the goal of identifying best practices to accelerate innovation in ophthalmic imaging and patient outcomes. This committee has also explicitly emphasized the need for standardization of DICOM to enable artificial intelligence approaches in eye disease research [48]. However, despite these calls for action, adoption remains low [47].

One of the barriers to DICOM utilization is the fragmented landscape of retinal imaging with different manufacturers and devices using a variety of formats. Ophthalmology manufacturers' adoption of DICOM remains disappointingly low, particularly compared to other imaging-centric specialties[49]. Current ophthalmic imaging devices use a wide range of formats including png, jpg, tiff, fda, as seen in widely used publicly available datasets. This inconsistency not only impedes clinical practice but also hinders research and collaborative efforts[50].

One of the main other reasons is the early development and utilization of proprietary file formats and imaging visualization software by major ophthalmic device vendors. These companies built vendor-specific ecosystems long before DICOM introduced ophthalmology specific standards. Vendors had already started heavily utilizing their proprietary systems in the 1990s and early 2000s. By the time the DICOM Standard committee released supplements tailored to ophthalmology, such as Supplement 90 for Ophthalmic Photography (2003), Supplement 110 for Ophthalmic Coherence Tomography (2007), and Supplement 197 for Ophthalmic Tomography Angiography (2017), proprietary formats were already extensively entrenched in ophthalmology practices. Since vendors had invested heavily in developing

their own software and workflows, there was less incentive to adopt a new, standardized format that would require substantial changes. A practical barrier to DICOM adoption is that achieving true standards compliance is difficult in real-world workflows. DICOM requirements are modality- and object-specific, distributed across extensive standards documentation, and commonly depend on conditional rules. Existing compliance tools are typically designed for manufacturers and engineers, may be challenging for research users to interpret, often require specialized training, and may not reflect the most current standards. As a result, research groups frequently lack practical tooling for (1) evaluating compliance in a transparent and actionable way and (2) converting exported files into standard-compliant DICOM objects without modifying underlying pixel data.

In this section, we present a retinal imaging DICOM standardization framework that consists of a compliance reporting tool that produces interpretable, standards-aligned summaries of file compliance, and a conversion workflow that generates NEMA-compliant DICOM objects while supporting study governance requirements such as de-identification. This framework was developed and applied in the context of a large, multimodal retinal imaging study (AI-READI)[51], demonstrating its practicality for preparing interoperable, research-ready imaging datasets (Figure 2.1.1)[52,53].

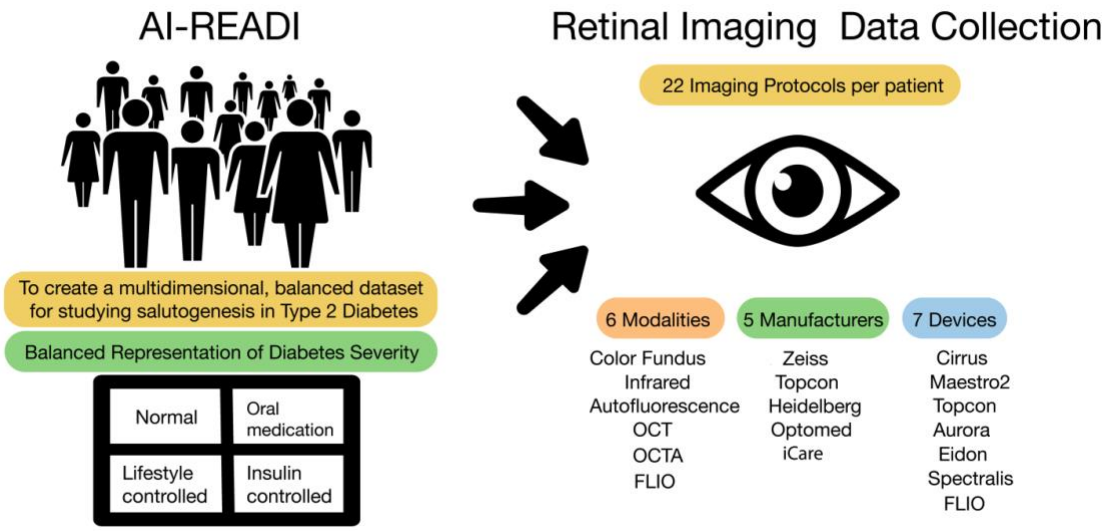


Figure 2.1.1. Summary of Artificial Intelligence Ready and Equitable Atlas for Diabetes Insights (AI-READI) retinal imaging data. (OCT: Optical Coherence Tomography; OCTA: Optical Coherence Tomography Angiography; FLIO: Fluorescence Lifetime Imaging Ophthalmoscopy). Reproduced from ref. 23 under a CC BY 4.0 license.

2.1.2 Methods

AI-READI Retinal Imaging Data collection

The retinal imaging data used to develop and evaluate the proposed standardization framework were obtained from the AI-READI project as part of the National Institutes of Health Common Fund's Bridge to Artificial Intelligence (Bridge2AI) program. AI-READI is a large, multimodal dataset designed to support AI-driven diabetes research and includes participants recruited from three academic medical

centers (University of Washington, University of California, San Diego, University of Alabama in Birmingham). The study population is designed to capture diversity in diabetes severity (no type 2 diabetes mellitus/normal, lifestyle-controlled, oral medication/non-insulin-controlled, insulin-controlled), race, and sex among adults aged 40 years or older[51,54].

The dataset includes retinal imaging data from six modalities, including color fundus photos (CFP), infrared (IR), fundus autofluorescence (FAF) images, optical coherence tomography (OCT), OCT angiography (OCTA), and fluorescence lifetime imaging ophthalmoscopy (FLIO) from five manufacturers including Topcon Healthcare (Oakland, NJ), Heidelberg Engineering (Franklin, MA), Carl Zeiss Meditec AG (Dublin, CA), Optomed USA (Alpharetta, GA), and iCare (Raleigh, NC). For each participant, approximately 44 retinal scans were acquired from both eyes across 6 modalities, 5 manufacturers, and 7 devices. Device exports consisted of heterogeneous file formats, including DICOM and non-DICOM formats (e.g., fda and proprietary vendor formats).

Standardization Workflow

We developed a multi-stage workflow to standardize heterogeneous retinal imaging exports into NEMA-compliant DICOM objects. The workflow consists of compliance evaluation, metadata correction and harmonization, and DICOM conversion (Figure 2.1.2). Files were first exported from the devices in their available formats, including DICOM, FDA, SDT, and HTML file types. For files already in DICOM format, NEMA compliance was evaluated using our compliance evaluator. Based on this evaluation, a DICOM conversion workflow was developed to correct tags and values to achieve NEMA compliance. The conversion tasks included ensuring standards compliance, incorporating missing metadata, harmonizing tag structures and values for standardization, and removing redundant information. Identifiable information was removed for the AI-READI study.

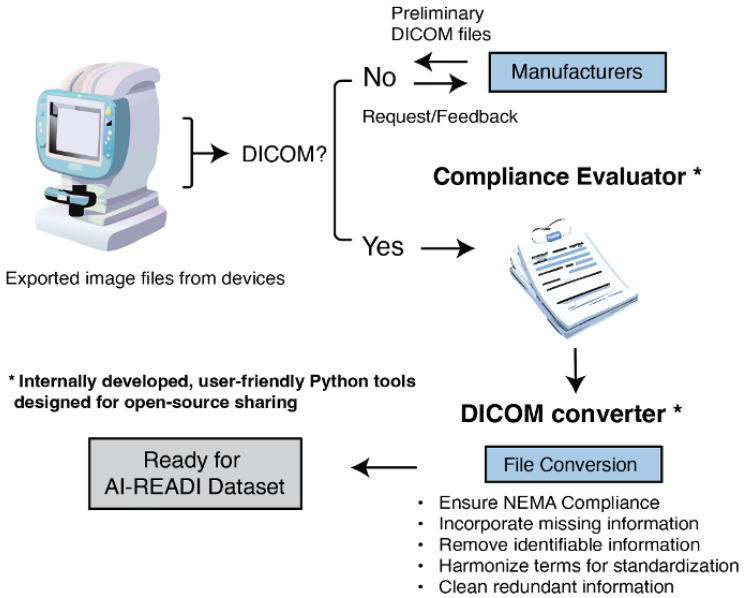


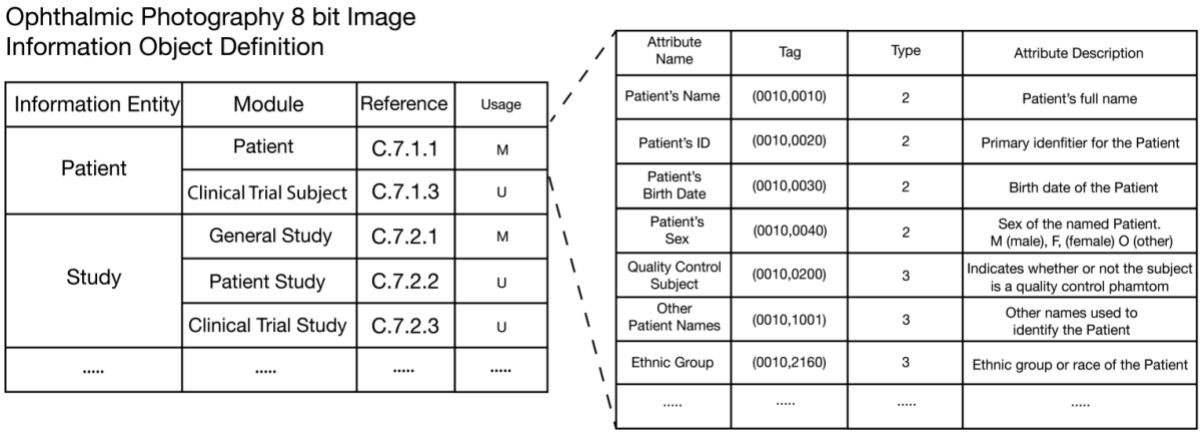
Figure 2.1.2. Overview of the imaging standardization workflow. Reproduced from ref. 52 under a CC BY-NC-ND 4.0 license.

For files not originally exported in DICOM format, manufacturer tools were requested to generate preliminary DICOM files. This step was necessary because many required DICOM attributes needed manufacturer validation to ensure accurate and preferred representations. The preliminary DICOM files were then evaluated using the compliance evaluator, and the results were communicated to the manufacturers. In case of missing information, the required details were subsequently provided by the manufacturers, and the DICOM converter was refined based on the validated metadata.

An exception was fluorescence lifetime imaging ophthalmoscopy (FLIO), for which no dedicated DICOM standard currently exists. For this modality, we developed a provisional DICOM representation by defining a structured mapping from FLIO-specific metadata into existing DICOM constructs. This approach enabled standardized storage and interoperability while preserving modality-specific information and maintaining compatibility with the broader standardization pipeline.

Compliance report generator

The standardization pipeline follows the most recent ophthalmic imaging specifications defined by the DICOM Standards Committee Working Group 09 (WG-09: Ophthalmology). The compliance report generator evaluates files according to five ophthalmic Service-Object Pair (SOP) Classes: (1) Ophthalmic Tomography Image Storage, (2) Ophthalmic Photography 8-Bit Image Storage, (3) B-scan Volume Analysis Storage, (4) Surface Segmentation Storage, and (5) En Face Image Storage. Color fundus photography, infrared, and autofluorescence scans correspond to SOP Class (2). OCT scans generate two file types corresponding to SOP Classes (1) and (2), while OCTA scans produce up to five file types spanning SOP Classes (1)-(5), with the number of en face images depending on anatomical location.



Ophthalmic Photography 8 bit Image
Information Object Definition

Information Entity	Module	Reference	Usage
Patient	Patient	C.7.1.1	M
	Clinical Trial Subject	C.7.1.3	U
Study	General Study	C.7.2.1	M
	Patient Study	C.7.2.2	U
	Clinical Trial Study	C.7.2.3	U
.....	.....	.....	.....

Attribute Name	Tag	Type	Attribute Description
Patient's Name	(0010,0010)	2	Patient's full name
Patient's ID	(0010,0020)	2	Primary idenfitier for the Patient
Patient's Birth Date	(0010,0030)	2	Birth date of the Patient
Patient's Sex	(0010,0040)	2	Sex of the named Patient. M (male), F, (female) O (other)
Quality Control Subject	(0010,0200)	3	Indicates whether or not the subject is a quality control phamtom
Other Patient Names	(0010,1001)	3	Other names used to identify the Patient
Ethnic Group	(0010,2160)	3	Ethnic group or race of the Patient
.....	.....	.....	.....

Figure 2.1.3. Example DICOM standards module from DICOM PS3.3 standards for ophthalmic photography 8-bit Image Information Object Definition (IOD). Reproduced from ref. 23 under a CC BY 4.0 license.

Each SOP Class contains approximately 70-200 defined tags with requirement categories specified by the National Electrical Manufacturers Association (NEMA). Requirement types include Type 1 (tag and value required), Type 2 (tag required, value may be empty), Type 3 (optional tag and value), and conditional types (2C/3C), where requirements apply only under predefined conditions. Separate compliance evaluation modules were implemented for each SOP Class to account for their distinct rule sets. For example, the Ophthalmic Photography 8-Bit Image IOD modules (Figure 2.1.3) include the

information entity, Patient, which lists its modules Patient, and Clinical Trial Subject with their respective references (C.7.1.1 and C.7.1.3), and whether the module is optional (U) or necessary (M). The reference for the necessary Patient module (C.7.1.1), lists associated tags and related information, such as the tag name, tag code, tag type, and the description of the value (attribute). Tag types are organized as 1 (tag and value required), 2 (tag required but value can be empty), 3 (tag and value are optional), and C (conditional - tag and/or value are required only under specific, predefined conditions).

The compliance report generator was developed based on the NEMA standards. Given a file or list of files, it generated a compliance report in a table format (e.g. xlsx) where the user can view all tag information from multiple files at a glance, eliminating the need to click on different links to view required compliance information for each SOP class. The table was an extension of each SOP class IOD Modules to a tag level. The table was color coded based on the compliance level. For example, if the tag needed both tag and value (type 1) but they were both missing the cell was marked red, for tags that had tag but no value when both are needed, they were marked yellow, and for optional values that were empty, they were marked blue. The purpose of this tool is to evaluate the NEMA compliance of the file content at a glance in a user-friendly manner. For file comparisons, multiple files could be viewed together as the table will become longer horizontally with more files (Figure 2.1.4).

Information Entity	Module	Reference	Tag	Element Name	Example DICOM File 1	Example DICOM File2
Study	General Study	C.7.2.1	00080020	Study Date	20230221	20220328
			00080030	Study Time	134639	142837
			00081030	Study Description	Optional	Optional
Frame of Reference	Synchronization	C.7.4.2	0018106A	Synchronization Trigger	NO TRIGGER	NO TRIGGER
Equipment	General Equipment	C.7.5.1	00080070	Manufacturer	VALUE NEEDED	UW
Image	Ophthalmic Photography Image	C.7.6.1	00080008	ImageType	ORIGINAL, PRIMARY	DERIVED, PRIMARY
			00280006	Planar Configuration	0	0
			00280030	Pixel Spacing	TAG AND VALUE NEEDED	0.0063, 0.0063

Figure 2.1.4. Example compliance report. An example of a file that is not NEMA compliant, Example DICOM File 1 and a compliant file, Example DICOM File 2. Red is defined as a warning that both required tag and value information is missing. Yellow is defined as a warning that the required tag is present but required values are missing. Blue is defined as a warning that an optional tag is missing which can be left empty.

DICOM Converter

Using the compliance report as input, we implemented a device- and protocol-specific file-to-file DICOM conversion workflow based on relevant SOP Class UIDs (Table 2.1.1). This workflow operates on

DICOM exports and generates fully NEMA-compliant DICOM objects. The conversion performs operations including, incorporation of missing required metadata, harmonization of tag values and structures across devices, and removal of identifiable patient information according to study requirements. All metadata inspection and modification were implemented using the Python pydicom package[55], which enables programmatic access to DICOM data elements and supports reproducible, script-based standardization.

Manufacturer	Device	Manufacturer provided export tools for preliminary DICOM files	Reference SOP Class UIDs for converting preliminary DICOM files to NEMA compliant DICOM files
Optomed	OptoMed	Yes, JPEG to DICOM	1.2.840.10008.5.1.4.1.1.77.1.5.1
iCare	Eidon	N/A (Already exported as DICOM)	1.2.840.10008.5.1.4.1.1.77.1.5.1
Heidelberg	Spectralis	Yes, proprietary DICOM to DICOM	1.2.840.10008.5.1.4.1.1.77.1.5.1 1.2.840.10008.5.1.4.1.1.77.1.5.4 1.2.840.10008.5.1.4.1.1.77.1.5.8 1.2.840.10008.5.1.4.1.1.66.8 1.2.840.10008.5.1.4.1.1.77.1.5.7
Zeiss	Cirrus	Yes, proprietary DICOM to DICOM	1.2.840.10008.5.1.4.1.1.77.1.5.1 1.2.840.10008.5.1.4.1.1.77.1.5.4 1.2.840.10008.5.1.4.1.1.77.1.5.8 1.2.840.10008.5.1.4.1.1.66.8 1.2.840.10008.5.1.4.1.1.77.1.5.7
Topcon	Maesetro2	Yes, fda files to DICOM	1.2.840.10008.5.1.4.1.1.77.1.5.1 1.2.840.10008.5.1.4.1.1.77.1.5.4 1.2.840.10008.5.1.4.1.1.77.1.5.8 1.2.840.10008.5.1.4.1.1.66.8 1.2.840.10008.5.1.4.1.1.77.1.5.7
Topcon	Triton	Yes, fda files to DICOM	1.2.840.10008.5.1.4.1.1.77.1.5.1 1.2.840.10008.5.1.4.1.1.77.1.5.4 1.2.840.10008.5.1.4.1.1.77.1.5.8 1.2.840.10008.5.1.4.1.1.66.8 1.2.840.10008.5.1.4.1.1.77.1.5.7
Heidelberg	FLIO	No, .sdt and .html files were custom formatted by us to create DICOM	No formal rules as FLIO is a relatively new modality. We followed 1.2.840.10008.5.1.4.1.1.77.1.5.2 based on the DICOM working group's suggestion.

Table 2.1.1. Device-specific export pathways and reference SOP Class UIDs used in the retinal imaging DICOM standardization pipeline. The table summarizes manufacturer-provided tools used to generate preliminary DICOM files and the corresponding reference SOP Class Unique Identifiers (UIDs) applied during conversion to NEMA-compliant DICOM objects. For modalities lacking established DICOM standards (e.g., FLIO), a custom mapping approach was implemented based on guidance from the DICOM working group.

For imaging files originally exported in non-DICOM formats, manufacturers provided tools to generate preliminary DICOM versions. These preliminary exports were evaluated using the compliance tool and missing or inconsistent metadata were identified. Through an iterative process of feedback and validation with manufacturers, protocol-specific metadata mappings were refined. Our final conversion rules are

built on the most up-to-date manufacturer exports and perform complementary corrections to achieve standards compliance.

Because correct metadata varies by device, company, and imaging protocol, the conversion workflow is implemented as a set of protocol-aware, device-specific mappings. This iterative development process ensures accurate representation of protocol metadata and establishes a reproducible framework that can be extended to additional devices and protocols. Importantly, the conversion workflow does not modify pixel arrays and only metadata fields are altered, ensuring that downstream image analysis remains unaffected by the standardization process.

2.1.3 Results

Standardization of exported imaging files to NEMA compliant DICOM files

Using the compliance results, we developed a conversion workflow that produces NEMA-compliant DICOM objects while preserving the original pixel data. Across the included modalities and device exports, the workflow successfully generated standardized DICOM objects suitable for interoperable storage. At the dataset level, this framework enables consistent representation of heterogeneous retinal imaging exports into standardized DICOM objects.

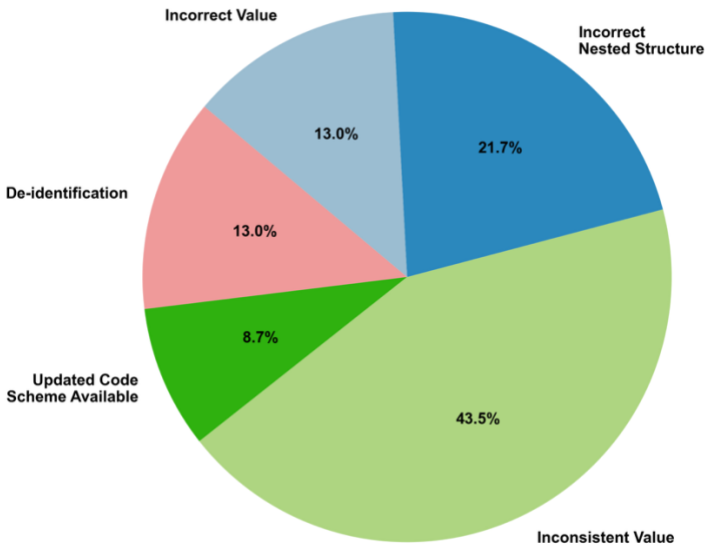


Figure 2.1.5. Categorization of Issues Requiring Harmonization in Ophthalmic DICOM data

A total of five file types, Ophthalmic Photography (OP), Ophthalmic Tomography (OCT), OCT B-scan Volume Analysis (Flow Cube), Segmentation, and OCT En Face (En Face) images, were evaluated based on their corresponding SOP Class UIDs and associated DICOM standards. Each file type's DICOM standards had a varying number of modules and tags associated with each module. A total of 113 unique tags were evaluated across the five types, with overlapping tags between types and nested structures counted as a single tag for consistency.

The standardization process included harmonization of tags/values (23 tags), and removal of demographic information (3 tags). For noncompliant Type 1 tags, which require both the tag and its associated values, the values were added either through the suggestion of a value, followed by confirmation from the manufacturer, or direct contribution from the manufacturer to ensure the information was correct. For noncompliant Type 2 tags, which require the tag but allow empty values, the tags were added during file conversion. In addition to adding missing tags and values, harmonizing existing values was necessary for reasons shown in Figure 2.1.5. The top reasons for harmonization, in decreasing order, were inconsistent values, incorrect nested structures, incorrect values, de-identification, and the availability of an updated code scheme.

Tags related to patient demographic information were harmonized to be empty or to have a unified value for de-identification purposes. Additionally, a tag was added to explicitly identify the imaging protocol.

The 2025 AI-READI retinal imaging dataset includes NEMA-compliant DICOM files collected and formatted across multiple vendors and modalities. Fundus and reflectance imaging includes 11,189 infrared reflectance scans from Heidelberg, 9,807 color fundus images from Optomed, 22,508 color fundus images, 4,506 infrared reflectance images from Topcon, and 18,343 infrared reflectance images from Zeiss. From iCare devices, the dataset includes 4,799 autofluorescence scans, 4,768 infrared reflectance scans, and 18,000 color photography scans. The structural OCT component includes 11,070 scans from Heidelberg, 18,343 from Zeiss, and 27,064 from Topcon, acquired across two Topcon devices (Maestro2 and Triton). The dataset further includes OCTA scans, consisting of 2,342 from Heidelberg, 13,386 from Topcon, and 8,832 from Zeiss. In addition, 7,968 fluorescence lifetime imaging ophthalmoscopy (FLIO) scans from Heidelberg are included. Altogether, across fundus photography, infrared reflectance, autofluorescence, structural OCT, OCTA, and FLIO modalities, the 2025 AI-READI dataset release comprises a comprehensive set of retinal imaging scans in NEMA compliant DICOM files.

2.1.4 Discussion & Conclusion

Our work proposes a practical blueprint for retinal imaging standardization that brings together multiple stakeholders, including manufacturers, the DICOM working group, and research users, and demonstrates implementation in a large real-world study. In a field that has historically lacked uniform standards adoption, this framework translates abstract DICOM specifications into an actionable workflow that spans the full pipeline from device export to NEMA-compliant DICOM objects. A key principle of the approach is manufacturer validation of standardized values, ensuring that protocol metadata are accurately represented while remaining interoperable.

A central contribution of this work is shifting standards compliance from a documentation-centric and manufacturer-oriented process into a transparent workflow usable by research teams. The compliance reporting tool enables interpretable, tag-level evaluation aligned with official standards, allowing users to identify and correct issues systematically. Its usability distinguishes it from existing tools that often present decontextualized warnings.

The conversion workflow complements the evaluator by performing protocol-aware, device-specific corrections. This component is necessarily more manufacturer-specific, because missing or ambiguous

metadata often cannot be inferred without vendor confirmation. As implemented, the workflow is directly applicable to the devices covered in the study and extensible through additional protocol mappings.

Several limitations reflect real-world constraints in imaging standardization. Some required metadata may be unavailable in device exports, and certain derived objects depend on proprietary vendor tooling. In these cases, the framework improves interoperability at the metadata and object-structure level, but full compliance may require updated manufacturer capabilities. Importantly, the workflow preserves original pixel arrays and modifies only metadata, preventing unintended bias in downstream analyses.

The standardization results demonstrate that achieving practical DICOM compliance requires systematic correction across multiple dimensions of metadata representation. Raw device exports frequently diverge from formal standards, and automated compliance workflows are therefore essential for constructing interoperable research datasets. Manufacturer collaboration enabled validation of protocol-specific attributes and established a reproducible process for extending the framework to future devices.

These technical corrections have broader implications for AI-driven research (Figure 2.1.6). Modern AI systems increasingly depend on large aggregated datasets to learn generalizable representations. Foundational ophthalmic models such as RetFound[56], OCTCube[57], and EyeFound[58] illustrate how performance scales with dataset diversity and size, yet heterogeneous file formats often require extensive preprocessing. Standardized DICOM representations reduce this burden by providing consistent metadata structure, facilitating scalable dataset integration and improving reliability in large-scale model training.

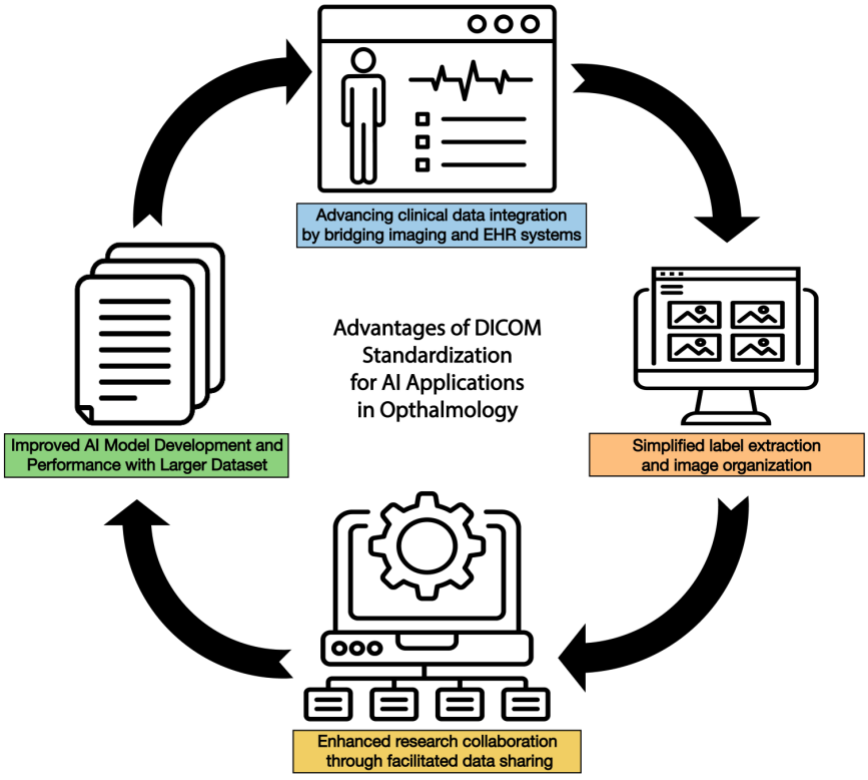


Figure 2.1.6. DICOM standardization for artificial intelligence applications in ophthalmology. (EHR: Electronic Health Records; AI: Artificial Intelligence). Reproduced from ref. 23 under a CC BY 4.0 license.

Standardization also supports tighter integration between imaging and structured health data. Associating DICOM metadata with electronic health record frameworks, including extensions of the OMOP Common Data Model[59], enables multimodal data-based analyses that combine imaging features with longitudinal clinical data. Such integration is essential for reproducible cohort formation and cross-institutional harmonization as multimodal systems become more prominent.

In addition, embedded DICOM metadata simplifies label creation and dataset organization. Demographic characteristics, acquisition parameters, and protocol descriptors can be automatically retrieved using reusable pipelines, reducing manual preprocessing and supporting consistent feature engineering. This is particularly important in ophthalmology, where demographic and device-specific factors influence disease characterization and model performance.

Additionally, DICOM is inherently designed to support data sharing through its robust network of communication standards. Services like DICOM Query/Retrieve and Storage Services enable efficient and secure sharing of files between systems. When integrated with networked PACS, DICOM facilitates real-time access to imaging data across multiple sites [60,61]. This capability is particularly useful for multi-institutional research projects, where joint efforts often depend on the rapid and reliable sharing of imaging data.

Overall, the proposed framework transforms heterogeneous retinal imaging exports into interoperable DICOM datasets suitable for governance-compliant research use. Beyond a purely technical formatting step, this work establishes standardization as enabling infrastructure for scalable AI development, multimodal medical research, and reproducible computational workflows in ophthalmic imaging.

2.2 Privacy Preserving Metadata-Driven Abstraction Layer for Tabular Health Data

2.2.1 Introduction

LLMs have potential to serve as a tool for translating natural language research questions into executable analytical workflows. However, direct interaction between LLMs and patient-level health data raises significant privacy, data governance, and security concerns[20–22]. Many high-performance commercial LLM systems operate through external commercial APIs, making it inappropriate or impermissible to expose raw clinical datasets to the model. At the same time, effective analytical automation requires the model to understand dataset structure, variable semantics, and relationships between tables.

This section introduces a privacy-preserving abstraction layer that separates schema understanding from data access. Instead of exposing patient records to the LLM, we construct a structured metadata representation, the Schema Summary Table, that captures the semantic meaning of variables using natural language descriptions derived from official documentation. This approach enables the LLM to reason about dataset structure and generates executable analytical code while maintaining strict isolation of sensitive patient information.

The metadata sharing step establishes a reusable interface between datasets and LLM API interactions, with all model communication restricted to the metadata layer. Together, these components enable scalable automation of hypothesis-driven research while preserving data privacy and auditability. The practical implementation of this interface within an end-to-end analytical framework is described in Chapter 4.

2.2.2 Methods

Prior to analysis, all datasets undergo a one-time preprocessing procedure designed to generate a comprehensive metadata artifact: the Schema Summary Table. This table enables the Large Language Model (LLM) to interpret the dataset's structure and generate necessary code to analyze them without any exposure to the underlying patient-level data. The procedure utilizes the dataset's publicly available data dictionary and documentation and consists of two primary stages (Figure 2.2.1).

First, in the code-to-text translation stage, the data dictionary is programmatically parsed to map technical variable identifiers (e.g., DIQ010) and their corresponding value encodings (e.g., 1) to their full natural language descriptions ("Doctor told you have diabetes") and labels ("Yes"). This step is necessary because column names and encoded values in structured datasets are often abbreviated, numeric, or based on dataset-specific coding systems that do not directly convey their clinical meaning without reference to external documentation. Translating these identifiers into explicit natural language descriptions ensures that the LLM can correctly understand the semantic intent of each variable rather than relying on ambiguous or underspecified column names.

Second, during the Schema Summary Table construction, metadata information is then consolidated and structured into a single reference table where columns include variable names, table names, natural language descriptions, data types, and example values, as available. The resulting Schema Summary Table serves as the sole source of information for the LLM during the subsequent query interpretation phase, thereby ensuring a strict separation between the model's semantic understanding of the data schema and the confidential patient data itself. All processing was conducted in Python using the Pandas[62] library.

a Code-to-Text Translation

b Schema Summary Table

Column Name	Column Description	Table Name	Table Description	Example Values
gender	What is the gender of the respondent?	demo_f	Demographic Variables	Female, Male
taking_insulin_now	Are you now taking insulin?	diq_f	Diabetes	Yes, No
...	...	...	...	...
albumin_g_dl	Albumin (g/dL)	biopro_f	Standard Biochemistry Profile	4.8, 3.0, 5.3

Figure 2.2.1. Overview of the data processing workflow. The Code-to-Text Translation step converts coded variable names into natural language descriptions using official metadata typically provided with the dataset. This enables the large language model (LLM) to understand variable meanings without prior knowledge of the coding system, thereby ensuring that the generated analytic code remains interpretable during human review. A Schema Summary Table is then created to summarize available variables and their attributes. During routine operation, only this table, no patient-level data, is shared with the LLM. This schema summary forms the basis for subsequent schema-grounding steps, allowing LATCH to map natural language concepts to corresponding database variables. Reproduced from ref. 22 under a CC BY 4.0 license.

NHANES

The data acquisition process involved the programmatic retrieval of all available XPT (SAS transport) files from the NHANES public data portal spanning 1999-2023, a total of 12 survey cycles across five major categories: Demographics, Dietary, Examination, Laboratory, and Questionnaire, which yielded 1,562 tables covering 128,809 unique participants.

Variable metadata and codebooks were obtained directly from the official NHANES documentation pages. For each dataset, corresponding codebook pages were scraped to extract definitions, Statistical Analysis System (SAS) labels, English-language descriptions, and value encodings. Parts of the NHANES ingestion pipeline were adapted by translating selected functions from the nhanesA[63] R package into Python. SAS natural language variable labels were selected as SQL column names because

they were both semantically descriptive and compatible with database's identifier length constraints. A minority of variables with identical codes but inconsistent SAS labels across years were resolved using a frequency-based harmonization strategy, where the most common label was adopted. This resulted in a single, unified metadata file that serves as the basis for the Schema Summary Table.

NHANES Mortality data

To enrich the dataset for survival analyses, associated mortality data were obtained from the NHANES Linked Mortality Files which are available for survey cycles 1999-2018 and had 101,316 unique participants. Mortality status, cause-of-death codes, and follow-up time were extracted from .dat files, which were converted into tables in csv files and were linked to NHANES data via the respondent sequence number (SEQN). Finally, all tables in csv files, including survey data and mortality files, were loaded into a database.

AI-READI Clinical data

The clinical data for the AI-READI Year 3 cohort[51] were obtained following approval through the standard AI-READI data access application process. The cohort consisted of 2,280 participants. The dataset was provided in the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) format and included the person, measurement, observation, and condition_occurrence tables. The procedure_occurrence table was not included in the final analytic dataset due to the presence of only one recorded procedure across the cohort, resulting in negligible informational value for feature generation. To create a unified, analysis-ready table from these files, a multi-step processing workflow was executed. All column names and variable labels were standardized to lowercase and replaced special characters with underscores, ensuring consistency for data joins and transformations. Concurrently, data quality was ensured by first removing duplicate rows based on the combination of person_id, description, and value, and then by excluding irrelevant or ambiguous survey items (e.g., timestamps, vague frequency questions) based on a predefined list.

Observational records were filtered and merged with the external OMOP concept dictionary to retrieve standardized human-readable labels associated with OMOP Concept IDs. A publicly available OMOP documentation table[64] was used to construct a concept mapping dictionary. To accommodate nested REDCap-style value dictionaries, the mapping procedure converted raw numeric and string-based encodings into standardized categorical representations, such as mapping binary indicators (for example, "1") to interpretable labels ("Yes"), thereby preserving semantic consistency across variables. Variables with excessively long descriptive labels were shortened to comply with character length limitations imposed by downstream reporting and analysis platforms. Additionally, variables with unclear or noninformative labels were renamed to improve interpretability. A structural transformation involved reshaping the dataset from the native OMOP long format, in which each row represents a single observation, into a wide-format representation. The observation and measurement tables were independently pivoted using person_id as the primary index, with standardized clinical variable labels serving as column headers.

AI-READI OCT Imaging Data

The retinal imaging data for the AI-READI Year 3 cohort[51,53] were obtained following approval through the standard AI-READI data access application process. Macular 6 × 6 mm structural optical coherence tomography (OCT) volumes were acquired using the Topcon Maestro2 imaging platform. For each scan, two corresponding Digital Imaging and Communications in Medicine (DICOM) files[65] were provided: a structural OCT volume and a segmentation (SEG) file containing multiple retinal boundary surfaces. In total, 4,506 paired OCT-SEG DICOM file sets were included in the dataset. File correspondence between OCT and SEG data was established using a study manifest table that included the relative paths to each associated pair. Segmentation and OCT volumes were read using pydicom[66], and only successfully paired files from a predefined subject list were analyzed. The segmentation pixel data were extracted as a 3-dimensional array with dimensions corresponding to retinal boundaries, B-scans, and A-scans.

Nine retinal layer boundary labels were obtained from the DICOM metadata from the tag Segment Sequence. For each adjacent boundary pair, layer thickness (in pixels) was computed at every A-scan and B-scan position as the absolute difference in depth between boundary coordinates. The resulting per-scan dataset included thickness measurements for all retinal layers in native pixel units. Physical scaling information was extracted directly from the OCT DICOM metadata using the PixelMeasuresSequence tag. Axial spacing, lateral spacing, and slice thickness were used to convert the derived pixel thicknesses into micrometers. For eyes with multiple scans, thickness maps were averaged at the eye level; subsequently, averages across both eyes were computed, when possible, to obtain per-person imaging features. The final dataset contains standardized retinal layer thickness values in micrometers, organized consistently by subject ID and eye laterality, and is suitable for downstream statistical modeling and multimodal integration.

2.2.3 Results

Application of the metadata-driven abstraction workflow produced unified and standardized schema representations across both large-scale public survey data and multimodal clinical cohorts. The resulting Schema Summary Table consolidated variable level metadata into a single structured reference artifact that preserved semantic descriptions, value encodings, and structural relationships while excluding all patient-level records. This abstraction layer enabled schema-level reasoning without exposing protected health information.

For the NHANES dataset, automated ingestion and harmonization of survey cycles spanning 1999-2023 resulted in a comprehensive metadata index comprising 59,867 variables across 1,562 tables and 128,809 unique participants over 12 survey cycles. Variable distribution reflected the breadth of the survey design, including 19,302 examination variables, 17,130 questionnaire variables, 14,345 laboratory variables, 7,573 dietary variables, 1,437 demographic variables, and 80 mortality-related variables. Programmatic codebook parsing resolved inconsistencies in variable naming and labeling across cycles through a frequency-based harmonization strategy, resulting in a unified schema suitable for longitudinal and cross-cycle analyses. Integration of linked mortality data further extended analytic capability to time-to-event and survival modeling within the same abstraction framework.

For the AI-READI Year 3 clinical cohort (N = 2,280 participants), transformation of OMOP-formatted tables into a standardized wide-format representation yielded 347 harmonized variables. These included 165 observation variables, 109 measurement variables, 31 condition_occurrence variables, 25 retinal layer

thickness imaging features, 17 variables regarding participants and person tables. Integration with the OMOP concept dictionary ensured consistent mapping of concept identifiers to standardized human-readable clinical terminology. Data cleaning, deduplication, and value harmonization procedures improved semantic consistency and interpretability, enabling downstream analytical modeling and natural language schema grounding. Derived variables from the retinal imaging processing step were successfully incorporated into the same metadata abstraction layer used for tabular clinical data, demonstrating the extensibility of the framework to multimodal data types.

Across datasets, the Schema Summary Table functioned as the exclusive interface between raw data and the LLM. All model interactions were restricted to metadata artifacts, ensuring that analytical planning, schema-based reasoning, and relevant code generation occurred without direct exposure to patient-level records. These results demonstrate the feasibility of a privacy-preserving metadata abstraction architecture capable of supporting automated, hypothesis-driven research across heterogeneous health data sources.

2.2.4 Discussion & Conclusion

This section demonstrates that privacy-preserving LLM interaction with health data is achievable through metadata abstraction rather than direct data exposure. By separating semantic schema understanding from patient-level access, the proposed framework enables automated analytical reasoning while maintaining strict governance boundaries.

A key contribution of this work is the formalization of the Schema Summary Table as a reusable abstraction layer. Many existing approaches to AI-assisted data analysis implicitly assume that models have unrestricted access to datasets. In contrast, this design treats metadata as a controlled interface that encodes dataset semantics in natural language while shielding sensitive records. This architecture supports deployment in regulated research environments.

Several limitations should be acknowledged. Metadata quality is dependent on the completeness and accuracy of source documentation. Inconsistent or ambiguous codebooks may require manual curation to ensure that they have sufficient information for schema interpretation. Additionally, the abstraction layer focuses on structured and semi-structured tabular representations, and unstructured clinical text and raw imaging interpretation fall outside the scope of the current design.

Despite these limitations, the proposed architecture establishes a foundational layer for secure, scalable LLM-assisted analytics. By enabling models to reason about dataset structure by utilizing column, table names, descriptions, and example values, without accessing protected records, this framework reduces privacy risks while preserving analytical flexibility. It provides the necessary interface between standardized health data and the automated hypothesis-testing system introduced in the next chapter.

In summary, the metadata-driven abstraction layer functions as both a governance mechanism and a technical enabler. It transforms clinical datasets into interpretable schema representations that support automated research workflows while maintaining strict separation between semantic reasoning and sensitive data. This design is a prerequisite for the LATCH framework presented in Chapter 4 and for broader adoption of AI-assisted clinical research infrastructure.

2.3 Chapter Summary

This chapter established the data and governance foundations necessary for LLM-assisted clinical research. First, it presented a retinal imaging DICOM standardization framework that transforms heterogeneous ophthalmic imaging exports into interoperable, NEMA-compliant DICOM objects, which enables reproducible and governance-compliant use of multimodal imaging data. Second, it introduced a privacy-preserving metadata abstraction layer for tabular health data, in which schema-level information is translated into natural language representations that can be shared with an LLM without exposing patient-level data. Together, these contributions provide the standardized and privacy-aware infrastructure required for scalable automated analytics and form the technical basis for the LATCH framework introduced in Chapter 4 and 5.

Chapter 3. Manual Clinical Research Workflows: A Case Study Using the IRIS Registry

Before clinical research workflows can be meaningfully automated, it is necessary to understand how they operate in practice. Traditional hypothesis-driven clinical research involves a sequence of manual steps, including cohort definition, data extraction, cleaning, harmonization, and statistical analysis. These steps are often repeated across studies and datasets, requiring substantial domain expertise and programming effort. Despite advances in computational tools, much of this process remains labor-intensive and difficult to scale, limiting reproducibility and slowing the pace of evidence generation. This chapter examines a real-world example of a manual clinical research pipeline to characterize its structure, complexity, and limitations.

We present a case study investigating the relationship between post-intraocular pressure (IOP) elevation and the development of primary open-angle glaucoma (POAG) using data from the IRIS (Intelligent Research in Sight) Registry. Beyond its clinical objectives, this study serves as a model example for analyzing how traditional research workflows are constructed and executed. By documenting each stage of the manual pipeline and evaluating the associated effort, bottlenecks, and repetitive workflows, this chapter identifies patterns that constrain scalability. These findings motivate the need for structured automation frameworks and inform the design principles of the LLM-assisted system introduced in subsequent chapters. Portions of this chapter are adapted from previously published work[78] by the author.

3.1 Clinical Significance of Post-IOP Elevation and POAG Development

3.1.1 Introduction

In adults aged 50 years and older, cataract is the most common cause of blindness in the world, followed by glaucoma.[67–69] Cataracts are a reversible cause of blindness, and cataract surgery is the most commonly-performed surgical procedure of all medical specialties, with an estimated 20 million per year worldwide.[70,71] Vision loss from glaucoma, on the other hand, is irreversible.[72] The most common type of glaucoma is POAG, comprising 60-80% of all glaucoma.[73–75] Most glaucoma progresses

asymptomatically until late stages, when advanced vision loss occurs.[76,77] Therefore, early detection and treatment of glaucoma is of utmost importance.[78]

Previous studies have focused on the effect of cataract surgery on short- and long-term IOP in patients with glaucoma. Postoperatively, IOP can be elevated, particularly among patients with pre-existing glaucoma.[79–85] However, these IOP elevations can also occur in patients without pre-existing glaucoma or other optic nerve pathology. These transient IOP elevations in patients without glaucoma have not yet been definitively investigated but are not believed to cause permanent damage.[86] In the long term, cataract surgery has been reported to be beneficial in reducing IOP, particularly in patients with higher preoperative IOP or angle-closure glaucoma.[83,87–94]

Understanding the relationship between short-term postoperative IOP elevation after cataract surgery and risk of glaucoma development in patients without prior glaucoma may have intraoperative or postoperative management implications. Therefore, we examined associations between postoperative IOP following cataract surgery and incident POAG and any glaucoma risk among 1,912,101 adults without prior glaucoma or ocular hypertension using data from the American Academy of Ophthalmology's IRIS® Registry (Intelligent Research in Sight)[95]

3.1.2 Methods

This study was conducted in accordance with the Declaration of Helsinki. Given the use of deidentified patient data, this study was exempted from review by the University of Washington Institutional Review Board. Data collection and aggregation methods used for the IRIS Registry database have previously been described.[96–99] Version chicago_amc_2023_04_14 of the IRIS Registry was used for this analysis. The IRIS Registry began collecting data in 2013, but some past ocular history is included in the database. Data were queried from the IRIS Registry using PostgreSQL 8.0.2 on Redshift 1.0.46987, and software used for data analysis included Python 3.8.5 and R 4.3.1.

Patient Selection

From the IRIS Registry, individuals aged 40 years or older who underwent standard extracapsular cataract extraction removal with insertion of intraocular lens from 2013 to 2023 were included using Current Procedural Terminology (CPT) code 66984, while patients who underwent complex cataract surgeries (CPT code 66982), multiple cataract surgeries, cataract surgery on eyes with unknown laterality, or additional eye procedures on the same surgery date were excluded. Individuals were excluded if they had a diagnosis of glaucoma or glaucoma suspect status or documented ocular hypertension prior to cataract surgery. Glaucoma-related diagnoses were identified using International Classification of Disease (ICD), Ninth Revision (ICD-9), and Tenth Revision (ICD-10) codes.

Glaucoma diagnoses included open-angle glaucoma, angle-closure glaucoma, secondary glaucoma due to medications, trauma, other eye disorders or inflammation, and unspecified glaucoma. Individuals with ocular hypertension, identified by any instance of IOP over 21 mmHg prior to cataract surgery, were

excluded. Patients were categorized as having commercial insurance if they had any commercial insurance coverage, regardless of government coverage.

Variables of Interest

The primary outcome of interest was the incident diagnosis of POAG, while the secondary analyses considered any glaucoma diagnosis. To identify patients who developed POAG after cataract surgery, we used ICD codes (365.1X and H40.11X). For identifying any type of glaucoma, including both primary and secondary forms, we used a broader set of ICD codes. If the initial diagnosis included POAG along with another type of glaucoma with a known cause (e.g., drug-induced, trauma-related) or a non-specific type (e.g., other specified glaucoma, unspecified), we did not categorized as POAG. This ensured that the only explicit cases of POAG were treated as such. To reduce potential bias due to the disproportionate representation of certain subgroups of a covariate during extended follow-up periods, right censoring of the data at 4,000 days was implemented. Variables included age, sex, race, ethnicity, and postoperative IOP.

The age at cataract surgery was calculated by subtracting the year of birth from the year of the cataract surgery. For the age analyses, patients were grouped into 10-year age groups. “Other” and “Unknown” for race in the IRIS Registry were excluded from the demographic descriptions and analytic dataset. Regarding “Sex”, “Race”, and “Ethnicity” in the IRIS Registry database, these data may be documented differently across practices.

The highest intraocular pressure recorded on postoperative days 0-2 was used for analyses. For Kaplan-Meier survival estimates, Cox proportional hazards model, and stratified Cox model analysis, the IOP was dichotomized, with values > 21 mmHg designated as “high” and values ≤ 21 mmHg designated as “normal”. For the more granular analysis of IOP level on risk of glaucoma, IOP was categorized into decile categories (e.g., ≤12 mmHg [0-10%]), and hazard ratios were estimated for each category, with the 40-60% IOP range as the reference, adjusting for demographic factors.

Statistical Analyses

Kaplan-Meier survival curves were used to estimate the cumulative probability of glaucoma diagnosis, including POAG only or any glaucoma, defined as the value of the survival probability subtracted from 1, across different subgroups within each demographic covariate. The log-rank test was applied to assess the time from cataract surgery to the onset of glaucoma diagnosis. Multivariable Cox proportional hazards analyses were conducted, providing hazard ratios (HRs) for age, sex, race, ethnicity, and postoperative IOP in relation to the risk of developing POAG and any glaucoma. Age was grouped in 10-year increments, with comparisons made to those who were 10 years younger. Reference groups were male for sex, White for race, non-Hispanic for ethnicity, and normal postoperative IOP. As a sensitivity analysis, multivariable Cox proportional hazards analyses were repeated after excluding patients who was diagnosed with POAG or any glaucoma diagnosis within 2 weeks following cataract surgery. This exclusion was implemented to reduce the possibility of misclassification, as patients may be transiently prescribed IOP-lowering medications after surgery.

For each covariate, multivariable Cox proportional hazards models were constructed within each subgroup of the covariate to explore the association between high postoperative IOP and risks of glaucoma, excluding the other subgroups as covariates. For example, risk of glaucoma in Hispanic or Latino patients with high postoperative IOP relative to Hispanic or Latino patients with normal postoperative IOP was calculated while controlling for the remaining covariates (i.e. age, sex, race, and IOP).

Additionally, for a granular analysis of IOP ranges on risk of glaucoma, we assessed risk using ten decile categories of the IOP distribution: ≤ 12, (12, 14], (14, 15], (15, 16], (16, 17], (17, 18], (18, 20], (20, 22], (22, 25], and > 25 mm Hg. Parentheses indicate greater than, and square brackets indicate less than or equal to. The IOP thresholds were determined based on the distribution observed among patients who did not develop glaucoma during our study. We used the Cox proportional hazard model to calculate HRs for each IOP category, adjusting for age, race, and ethnicity, with the middle decile groups, (16–18] mm Hg IOP range (40-60%] serving as the reference.

3.1.3 Results

Demographics

A total of 1,912,101 patients who underwent uncomplicated, unilateral cataract surgery in one eye without a history of glaucoma, ocular hypertension, or glaucoma suspect were identified in the IRIS Registry. The mean and standard deviation of the age of the cohort was 71 ± 8.5 years, and 1,145,854 (59.9%) patients were female. A total of 1,504,633 (78.7%) patients had normal postoperative IOP (≤21 mmHg) and 407,468 (21.3%) patients had high postoperative IOP (>21mmHg). During the follow-up period of 4,000 days, 10,710 (0.56%) were diagnosed with POAG. The median time to development of POAG was 682 days (IQR 191 - 1467 days).

Kaplan-Meier analysis and risk of POAG

Long-term follow-up with KM estimates showed that the 4,000-day cumulative probability of POAG diagnosis was 3.4% for patients with high postoperative IOP (>21 mmHg), nearly twice the 1.7% observed in the normal postoperative IOP group (≤21 mmHg) (log-rank test $P < 0.0001$) (Figure 3.1.1A).

Among racial groups, Black patients had the highest cumulative probability of diagnosis of POAG (5.0%), followed by Asians, Native Hawaiian and Other Pacific Islanders, and American Indians or Alaska Natives. White patients had the lowest percentage (1.9%). Regarding ethnicity, Hispanic or Latino patients were associated with higher cumulative probabilities (3.0%) compared to non-Hispanic or Latino patients (2.1%). Older age was associated with a higher cumulative probability of POAG diagnosis. Male sex was associated with higher cumulative probabilities (2.6%) than female sex (1.8%).

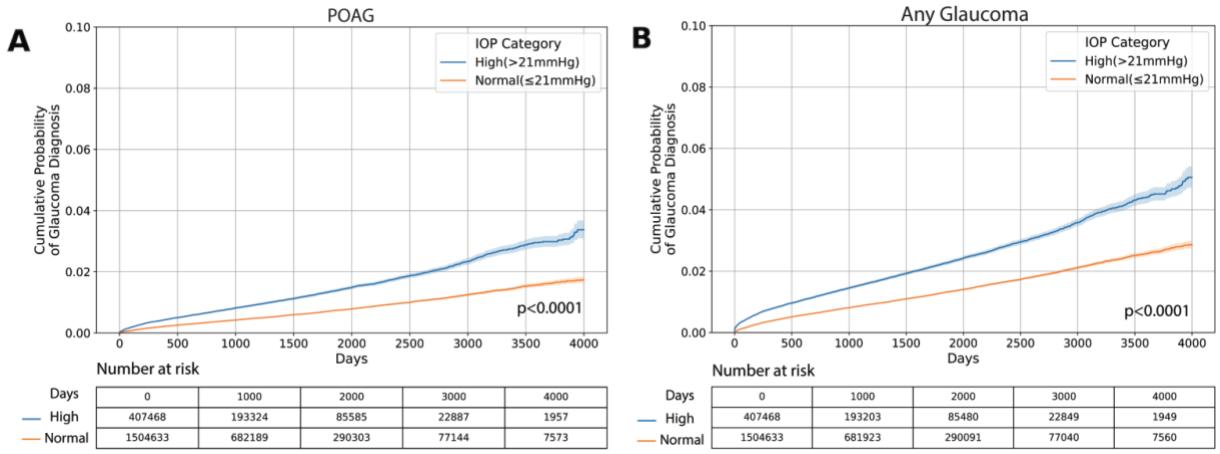


Figure 3.1.1. Kaplan-Meier survival curves showing the cumulative probability of A) POAG and B) any glaucoma following cataract surgery, stratified by IOP. The number of subjects at risk at days 0, 1000, 2000, 3000, and 4000 after cataract surgery are provided. The P-values from the log-rank tests were displayed. Reproduced from ref. 78 under a CC BY-NC-ND 4.0 license.

Cox Proportional Hazard Analysis of the association between postoperative IOP and risk of POAG

In the Cox proportional hazards analysis, high postoperative IOP was a significant risk factor for POAG (HR = 1.87, 95% CI: 1.79-1.94, P < 0.001) (Figure 3.1.2A). Each decade of increased age was associated with an increased risk of POAG (HR = 1.18, 95% CI: 1.15-1.20, P< 0.001). Male sex, as well as Asian, Black, Native Hawaiian and Other Pacific Islander races, along with Hispanic or Latino ethnicity, were associated with a higher risk of developing POAG. Notably, Black patients were nearly three times as likely to develop glaucoma compared to White patients, after adjusting for other covariates (HR = 3.37, 95% CI: 3.18-3.56, P < 0.001). In the sensitivity analysis, which excluded 533 individuals diagnosed with POAG within 14 days of cataract surgery, high IOP remained a significant risk factor for POAG (HR 1.82, 95% CI: 1.75-1.90, P < 0.001).

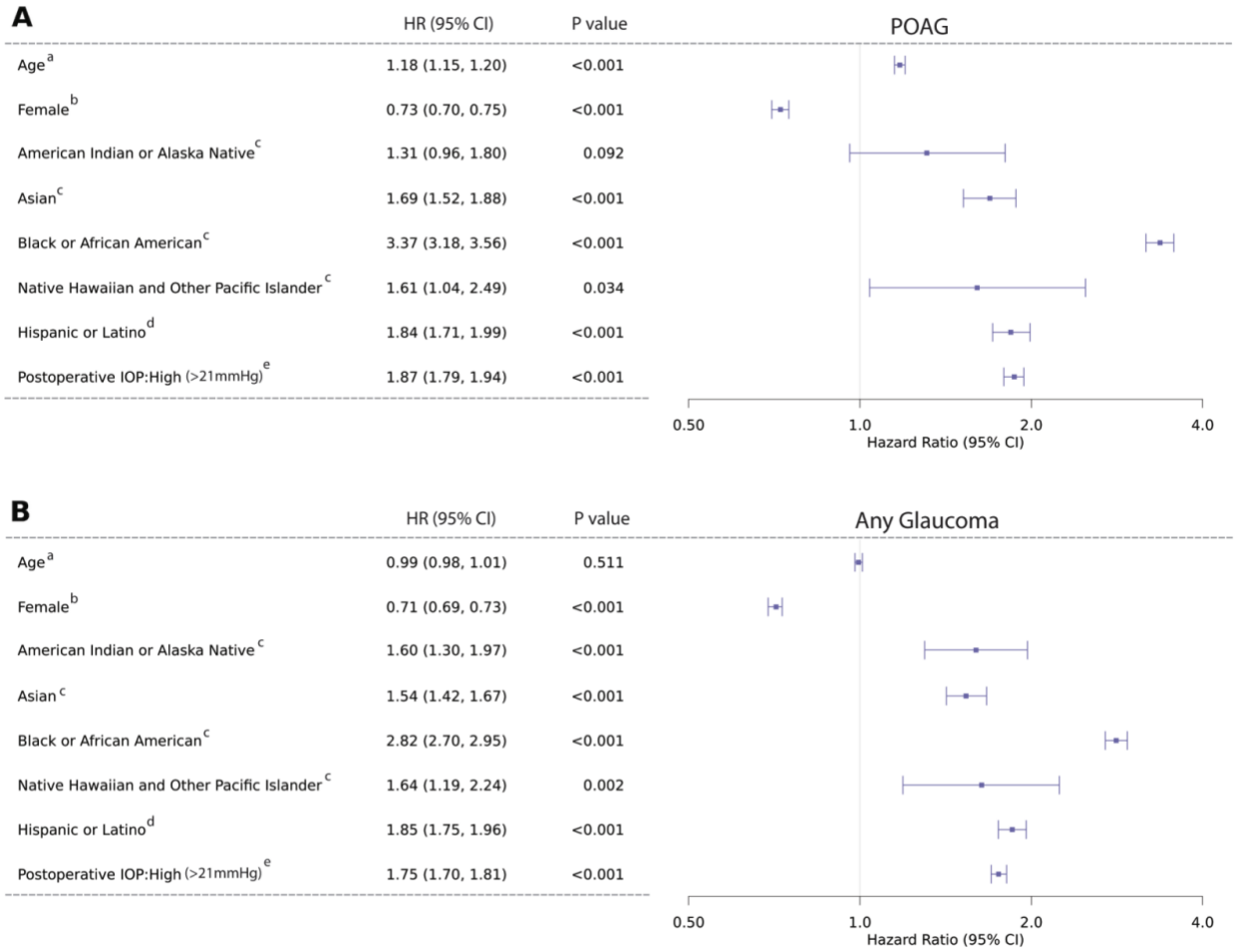


Figure 3.1.2. Multivariable Cox proportional-hazards analyses of demographic variable assessing postoperative IOP and risk of A) POAG and B) any glaucoma. Forest plots, presented on a log scale, illustrate the hazard ratios (HRs) for glaucoma, with 95% confidence intervals (CIs). [a]Age was grouped into decades of life and compared to the age group 10 years younger. Reference groups: [b]Male, [c]White, [d]Non-Hispanic/Latino, and [e]Normal postoperative IOP (≤21 mmHg). Reproduced from ref. 78 under a CC BY-NC-ND 4.0 license.

Stratified Cox proportional hazard analysis of risk of POAG by postoperative IOP in demographic subgroups

The stratified Cox proportional hazards analysis consistently identified high postoperative IOP as a significant predictor of POAG following cataract surgery, with high hazard ratios exceeding 1.5 across all demographic subgroups when compared to individuals with normal postoperative IOP. (Figure 3.1.3A). The magnitude of this association varied by demographic subgroup. By age, the effect of high postoperative IOP on POAG was more noticeable in patients aged 80 years and more, who had HR of 2.02 (95% CI: 1.84-2.21, P < 0.001) compared to those with normal postoperative IOP. Patients aged 40-79 had a slightly lower, though still elevated, HR of 1.83 (95% CI: 1.75-1.92, P < 0.001). When stratified by sex, female participants with high postoperative IOP had a slightly greater risk of developing POAG (HR = 1.90, 95% CI: 1.79-2.01, P < 0.001) than male participants (HR = 1.84, 95% CI: 1.73-1.94, P < 0.001), relative to individuals of the same sex with normal IOP. When stratified by ethnicity, non-Hispanic or Latino participants showed a stronger association between high postoperative IOP and POAG

(HR = 1.89, 95% CI: 1.82-1.97, P < 0.001) compared to Hispanic or Latino participants (HR = 1.51, 95% CI: 1.28-1.77, P < 0.001), demonstrating a less pronounced effect in the Hispanic subgroup. Subgroup analyses based on race showed an elevated risk in Asian participants with high postoperative IOP (HR = 1.98, 95% CI: 1.59-2.48, P < 0.001), followed by White (HR = 1.90, 95% CI: 1.82-1.99, P <0.001) and Black or African American (HR = 1.64, 95% CI: 1.47-1.83, P <0.001) participants, each compared to participants in the same race group with normal IOP. Hazard ratios were also elevated for American Indian or Alaska Native and Native Hawaiian (HR = 2.33, 95% CI: 1.21-4.50, P =0.012) or Native Hawaiian and Other Pacific Islander groups (HR = 2.01, 95% CI: 0.77-5.25, P =0.156), although with lesser level of statistical significance. Despite all groups demonstrating increased risk, there was variability in the degree of association between high postoperative IOP and POAG.

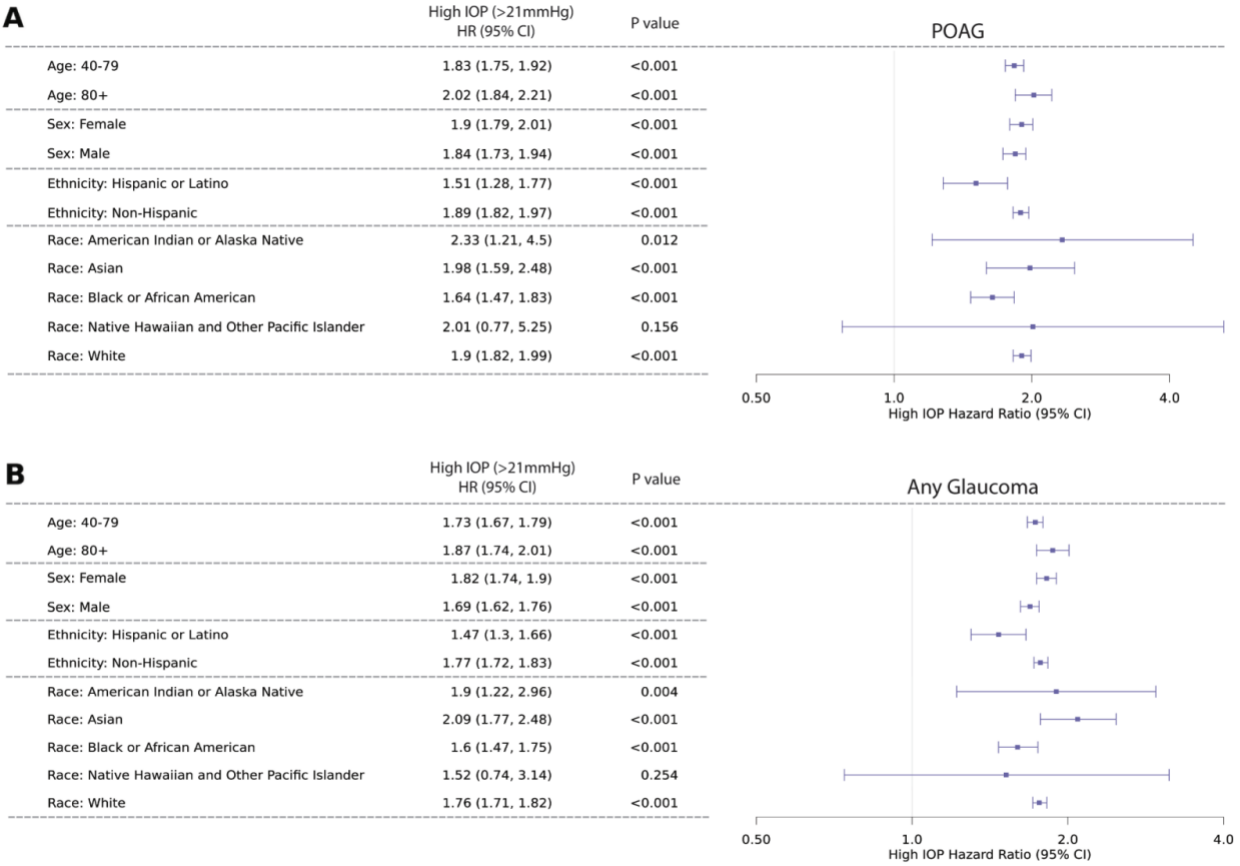
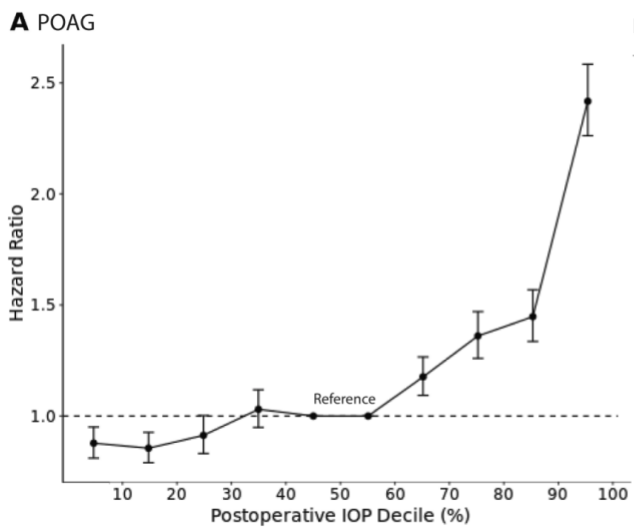


Figure 3.1.3. Stratified Cox proportional hazard analysis of risk of POAG by postoperative IOP in demographic subgroups. Adjusted hazard ratios (HRs) is shown for A) POAG and B) any glaucoma, associated with high postoperative IOP. Forest plots, presented on a log scale, display adjusted HRs for glaucoma in the high IOP group compared to the normal IOP group within a specific subgroup. Horizontal lines represent 95% confidence intervals (CIs) for each HR, showing the variability in the association between high IOP and glaucoma across subgroups. Reproduced from ref. 78 under a CC BY-NC-ND 4.0 license.

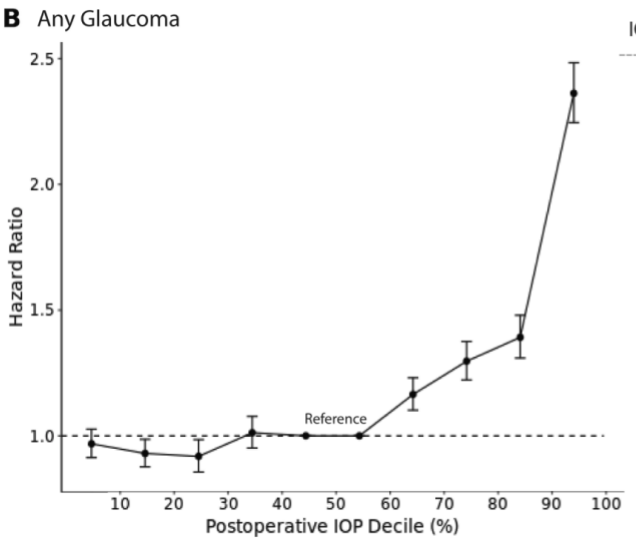
Risk Assessments of POAG Across Deciles of Postoperative IOP Levels

The top decile postoperative IOP was associated with a HR of 2.42 (95% CI: 2.26-2.58, P < 0.001) for POAG, compared to the middle reference deciles (Figure 3.1.4A). We observed a gradual increase in HRs

of glaucoma risk across the increasing IOP deciles for POAG. Patients in the 0-10% and 10-20% deciles of the postoperative IOP distribution had a significantly lower risk of POAG than the middle reference deciles. For example, the 0-10% group had an HR of 0.88 (95% CI: 0.81-0.95, P = 0.001), and the 10 - 20% group had an HR of 0.86, 95% CI: 0.79-0.93, P < 0.001).



IOP Decile	IOP range (mmHg)	HR (95% CI)	P value
0-10%	≤ 12	0.88, (0.81, 0.95)	0.001
10-20%	(12, 14]	0.86, (0.79, 0.93)	<0.001
20-30%	(14, 15]	0.91, (0.83, 1.00)	0.056
30-40%	(15, 16]	1.03, (0.95, 1.12)	0.484
40-60%	(16, 18]	1.00, (- , -)	-
60-70%	(18, 20]	1.18, (1.09, 1.27)	<0.001
70-80%	(20, 22]	1.36, (1.26, 1.47)	<0.001
80-90%	(22, 25]	1.45, (1.34, 1.57)	<0.001
90-100%	> 25	2.42, (2.26, 2.58)	<0.001



IOP Decile	IOP range (mmHg)	HR (95% CI)	P value
0-10%	≤ 12	0.97, (0.91, 1.03)	0.278
10-20%	(12, 14]	0.93, (0.88, 0.99)	0.015
20-30%	(14, 15]	0.92, (0.86, 0.99)	0.017
30-40%	(15, 16]	1.01, (0.95, 1.08)	0.699
40-60%	(16, 18]	1.00, (- , -)	-
60-70%	(18, 20]	1.16, (1.10, 1.23)	<0.001
70-80%	(20, 22]	1.30, (1.22, 1.38)	<0.001
80-90%	(22, 25]	1.39, (1.31, 1.48)	<0.001
90-100%	> 25	2.36, (2.25, 2.48)	<0.001

Figure 3.1.4. Hazard ratios (HRs) of postoperative IOP on A) POAG and B) any glaucoma by IOP deciles (0-10%, 10-20%, 20-30%, 30-40%, 40-50%, 50-60%, 60-70%, 70-80%, 80-90%, 90-100%). The y axis indicates hazard ratios (HRs) with error bars representing 95% confidence intervals (CIs) for each IOP category, relatively to the 40–60% IOP category, which serves as the reference. Deciles correspond to ≤ 12, (12, 14], (14, 15], (15, 16], (16, 17], (17, 18], (18, 20], (20, 22], (22, 25], and > 25 mm Hg. Parentheses indicate greater than, and square brackets indicate less than or equal to. Reproduced from ref. 78 under a CC BY-NC-ND 4.0 license.

Secondary analysis of postoperative IOP and risk of any glaucoma

The same set of analyses was performed with any incident glaucoma as the outcome instead of POAG, as a secondary analysis. The 4,000-day cumulative probability of any glaucoma diagnosis was 5.1% for the

high postoperative IOP group and 2.9% for the normal postoperative IOP group (Figure 3.1.1B). Race, ethnicity, and sex categories followed similar patterns for both POAG and any glaucoma, with American Indian or Alaska Native, Asian, Black, Native Hawaiian and Other Pacific Islander races, Hispanic ethnicity, and male sex associated with higher cumulative probabilities. However, what was different from the results for POAG was that older age was not associated with a higher cumulative probability of any glaucoma development.

In the Cox proportional hazards analysis, high postoperative IOP was a significant risk for any glaucoma (HR 1.75, 95% CI: 1.70-1.81, P < 0.001) (Figure 3.1.2B). Age was not associated with an increased risk of any glaucoma (HR = 0.99, 95% CI 0.98 - 1.01, P=0.5). The stratified Cox proportional hazards analysis consistently identified elevated postoperative IOP as a significant predictor of any glaucoma development, when compared to the normal postoperative IOP group within each subgroup (Figure 3.1.3B). Across all covariate subgroups, patients with elevated postoperative IOP had elevated HRs for any glaucoma compared to those with normal postoperative IOP. The highest HR was observed in the Asian group (HR = 2.09, 95% CI: 1.77-2.48, P <0.001). We observed a gradual increase in HRs of glaucoma diagnosis across the increasing IOP deciles for any glaucoma (Figure 3.1.4B).

3.1.4 Discussion & Conclusion

In this retrospective cohort study of 1,912,101 IRIS Registry participants with normal preoperative IOP and without a history of glaucoma, glaucoma suspect, or ocular hypertension diagnoses who underwent cataract surgery, high postoperative IOP nearly doubled the risk of developing POAG. Male sex, as well as Asian, Black, Native Hawaiian and Other Pacific Islander races, along with Hispanic or Latino ethnicity, were associated with increased risk of POAG, with Black patients showing nearly three times the risk compared to White patients. In the stratified Cox analysis, high postoperative IOP was consistently associated with an increased risk of POAG, with all HRs ranging from 1.51 to 2.33 when compared to the normal group within relevant subgroup. We observed a dose-dependent relationship between IOP elevation and risk of POAG. Similar patterns were observed with risks of any glaucoma.

Our study investigated the potential impact of postoperative IOP elevation in patients without a history of glaucoma, while previous studies have focused on patients with prior glaucoma. Patients with prior glaucoma have been shown to be more likely to have high IOP post cataract surgery.[87,88,100] Ahmed et al. reported that a significant number of patients with preexisting glaucoma had potentially harmful IOP spikes, defined as IOP > 28 mmHg, 3-7 hours after surgery, with a higher rate of elevated IOP in the glaucoma group (46%) compared to the non-glaucoma group (18%).[81] However, Ahmed et al. did not followed up patients beyond 4 days post-surgery, and did not have data on long-term risks of glaucoma. Another retrospective study with a longer follow-up of 30 days reported that a preoperative diagnosis of glaucoma was associated with a postoperative IOP higher than 22 mmHg after uncomplicated phacoemulsification.[79] Due to the increased prevalence of postoperative IOP spikes among patients with glaucoma or ocular hypertension, some support the use of prophylactic acetazolamide to lower IOP in this population when undergoing cataract surgery.[101] Despite this tendency toward IOP elevation in patients with glaucoma, elevated short-term postoperative IOP is often considered harmless in patients without glaucoma or other optic nerve pathology because the IOP usually decreases without intervention.[102]

There are several potential reasons for the development of glaucoma after cataract surgery in individuals without a history of ocular hypertension or glaucoma, especially in patients who experience early postoperative IOP elevation. One possibility is pre-existing subclinical aqueous outflow dysfunction, which may be revealed by surgical factors such as inflammation or ophthalmic viscosurgical devices (OVD). Outflow dysfunction may occur at different levels, such as trabecular, uveosleral, or the more recently proposed uveolymphatic systems.[103] However, these additional outflow pathways can be challenging to evaluate clinically, and dysfunction may go undetected prior to surgery. Additional mechanisms may involve intraoperative or postoperative components. For example, retained OVD[104–106] is a known cause of transient IOP elevation following phacoemulsification and may cause stress on the trabecular meshwork. Postoperative inflammation and prostaglandin release may also transiently increase IOP.[107] An IOP elevation from a steroid response may also occur postoperatively due to the use of ocular steroids in the management of inflammation post-surgery.

Whether transient IOP elevations secondary to retained OVD, postoperative inflammation, or management of said inflammation represent a predisposition to glaucoma versus causal relationships with glaucoma is unknown. However, given the nearly two-year interval before POAG diagnosis in our study population, our findings suggest a role for outflow dysfunction, while cannot definitively disentangle the contribution of cataract surgery itself to POAG risk. Our study was not designed to evaluate these surgical factors, as such details are not always shown in electronic health records, but their potential role adds to the complexity of understanding postoperative IOP changes. While our focus was on postoperative IOP elevation and future POAG risk, additional investigations may help determine whether this reflects pre-existing outflow vulnerability or is driven by surgical effects.

A recent IRIS Registry study by Lidder et al. investigated postoperative IOP spikes in stand-alone phacoemulsification.[82] They found that there was a greater risk of an IOP spike with glaucoma, higher baseline IOP, male sex, Black race, older age and complex surgery coding. In that study, glaucoma subtype affected the risk of an IOP spike; compared with glaucoma suspect eyes, those with secondary open-angle glaucoma (pigmentary and pseudoexfoliative subtypes), and ocular hypertension were more susceptible to an IOP spike. Eyes with primary angle closure (PAC) glaucoma, PAC suspects, and normal tension glaucoma were less susceptible to an IOP spike. It was theorized that eyes with secondary open-angle glaucoma may have difficulty achieving homeostasis following surgery due to a stressed trabecular meshwork, while those patients with PAC glaucoma and PAC suspects benefited from the increased space and therefore increased outflow with removal of the crystalline lens. This may support the theory that an IOP spike following cataract surgery in healthy eyes could signal a preexisting impairment of aqueous outflow that increases the risk of developing glaucoma.

While elevated IOP is a well-known risk factor for glaucoma, other significant risk factors include age, race and ethnicity, and genetic predisposition. We reported that Asian, Black, and Hispanic or Latino populations are at a higher risk for developing glaucoma[108,109], which was consistent with prior findings. Additionally, older age was associated with an increased risk of POAG in our study. Interestingly, we did not see a similar age-related risk increase for any glaucoma. This may have been due to the fact that the outcome for any glaucoma included diagnosis codes for glaucoma secondary to trauma and other glaucoma outcomes that are not associated with advanced age. This emphasizses the usefulness for subtype-specific analysis because different types of glaucoma may have distinct risk factors.

This study has several limitations. First, we excluded patients with any preoperative IOP measurement of >21 mmHg to rule out any patients with ocular hypertension who are at higher risk of developing glaucoma.[110] However, misclassifications are still possible if patients had transiently lower IOPs during these examinations. We included all IOP measurements prior to cataract surgery to try to mitigate this influence. Importantly, the median time to develop glaucoma in our study was 682 days (IQR 191-1,467 days), and this long interval between surgery and diagnosis suggests that the newly diagnosed cases were not likely to be glaucoma cases that were missed during the preoperative period. Instead, they are more likely to represent new-onset glaucoma outcomes.

Our sensitivity analysis, in which we excluded patients diagnosed with glaucoma within 14 days of surgery, accounted for cases likely related to short-term IOP management rather than real disease. The persistence of a significant association between high postoperative IOP and incident glaucoma risk in this analysis further supports our main findings. Of note, there is no clear standard definition for a clinically meaningful IOP spike, and comparison of prior studies may suffer from differences in its definition. Although there are differences in the risk for glaucoma between different documented racial and ethnic groups, it is important to recognize that race and ethnicity are social constructs confounded by factors such as social class, economic status, and nationality; therefore, these differences should not be mistaken as a genetic effect.

Additional limitations are related to data quality and completeness because electronic health records are subject to coding errors and missing information,[111–114] and ICD-9 and ICD-10 coding may not always fully capture the diagnosis of glaucoma due to potential coding inaccuracies (e.g., inconsistent or incomplete coding). Despite these limitations, ICD codes remain a commonly utilized approach for studying glaucoma, as demonstrated by several IRIS registry studies. [115,116] The IRIS Registry does not include all practices and includes a small percentage of academic medical centers. We did not have access to other confounders that may have contributed to varying risks of glaucoma. For example, additional metrics used in glaucoma diagnosis and IOP interpretation, such as visual field test results and pachymetry, were not available to the IRIS Registry analytic centers for our analyses.

The timing of IOP measurements, diurnal variations, and the use of different instruments may have introduced bias and variations. Patient medical history, comorbidities, surgical techniques, and post-operative care may have contributed to risks of glaucoma development but were not consistently available in a comprehensive manner for our analyses. Additionally, genetic predisposition is a factor in glaucoma development, but we did not have information about genetic risk in this population. Lastly, there is likely a lag between when the patient first develops glaucoma and when the first glaucoma diagnosis occurs. However, despite these factors, it is less likely that the near two-fold increase in glaucoma development and the dose-dependent increase in risk of glaucoma are due to random variation alone.

High postoperative IOP following cataract surgery was associated with a nearly double risk of glaucoma development among individuals without a prior glaucoma, glaucoma suspect, or ocular hypertension diagnoses, regardless of demographic factors. Such patients should be followed up more closely following cataract surgery for the future development of glaucoma.

3.2 Evaluating the Manual Workflow: Process Analysis

3.2.1 Introduction

While the clinical results in Section 3.1 demonstrate the scientific value of large-scale registry research, they do not capture the operational complexity required to produce such analyses. Clinical research workflows are often described abstractly as a sequence of cohort definition, data extraction, and statistical modeling. In practice, however, these steps involve extensive manual iteration and coordination that are rarely documented systematically.

To better understand the structure and effort associated with traditional workflows, we conducted a case study of the end-to-end research pipeline used to produce the IRIS Registry analysis in Chapter 3.1. Rather than evaluating clinical outcomes, this section focuses on the research process itself. By analyzing development artifacts and reconstructing workflow timelines, we aim to characterize how effort is distributed, where bottlenecks arise, and which components may benefit from automation.

The above analysis is intentionally framed as an in-depth single-case process study. Although limited in scope, the workflow examined here represents a realistic and complex example of hypothesis-driven clinical research using large observational datasets. Similar structural patterns are described in studies using other large-scale clinical databases, including CMS Medicare claims data[117], NHANES population surveys, and the MIMIC[118] critical care database, where research pipelines typically involve cohort construction from coded variables followed by the application of established conventional statistical methods.

To determine whether the observed effort distribution in hypothesis development, cohort construction, statistical analyses, and manuscript preparation reflects broader patterns rather than project-specific factors, we additionally conducted a focused, structured literature search of comparable IRIS Registry studies examining intraocular pressure outcomes following cataract surgery. This review assessed convergence in statistical methodology and contrasted the relative stability of analytical modeling with the variability of cohort construction workflows.

3.2.2 Methods

Workflow Stage Decomposition and Time Estimation

To quantify the effort associated with different components of the research pipeline, we estimated the time required for each of five workflow stages leading up to initial manuscript submission: hypothesis development/refinement, cohort specification, database query development (in SQL), statistical analysis (in R) and writing. Time estimates were reconstructed retrospectively using a combination of email communication records, and research updates. They were reviewed chronologically to identify periods of active work, major revision events, and transitions between workflow stages. These records served as temporal markers that allowed reconstruction of approximate working intervals associated with specific tasks.

Activities were mapped to workflow stages based on their primary function. For example, SQL query drafting and revision were assigned to database query development, while cohort definition discussions and inclusion/exclusion refinement were assigned to cohort specification. Email exchanges that documented problem-solving, clarification requests, or revision decisions were used to contextualize the purpose and duration of specific work intervals. Estimated duration within each stage was calculated as the relative proportion of total project effort. These estimates are approximate and intended to provide a comparative distribution of effort rather than precise time accounting.

Reusability Assessment

The reusability assessment focused exclusively on the technical components of the workflow associated with database query development and statistical analysis. We conducted an artifact review of SQL queries and analysis code in R generated during these stages to evaluate their potential applicability in future studies.

Each technical component was rated for reusability based on the degree of modification required for adaptation to a new but similar IRIS registry study with different predictors and outcomes. Components were classified as highly reusable if they could be applied with minimal modification (estimated <20% changes), moderately reusable if partial rewriting would be required (20–50% changes), and low reusability if substantial restructuring would be necessary (>50% changes).

Ratings were informed by the extent of study-specific logic identified in code. Conceptual and narrative stages of the workflow, such as hypothesis refinement, cohort specification, and manuscript writing, were excluded from this assessment, as they do not produce directly transferable technical artifacts. This evaluation provides a qualitative measure of how much of the computational workflow represents generalizable infrastructure versus study-specific implementation.

Cross-Study Convergence in Statistical Methodology

To contextualize the workflow findings beyond the single-case process analysis, we conducted a focused literature screen to characterize statistical methodology used in comparable IRIS Registry studies. A PubMed search was performed using the following query:

```
("Intelligent Research in Sight"[tiab] OR "IRIS"[tiab])
 AND ("Intraocular Pressure"[tiab] OR "IOP"[tiab])
 AND ("Cataract"[tiab] OR "Phacoemulsification"[tiab])
```

Filters were applied for publication dates between 2016 and 2026 (10-year window) and full-text availability. The search returned 372 records. After screening for actual usage of IRIS Registry and excluding the IRIS Registry study described in Section 3.1, eight eligible studies remained for methodological review.

For each study, statistical methods were extracted and categorized using predefined criteria: (1) descriptive reporting of baseline characteristics, (2) group-stratified reporting and/or comparative statistical analyses, (3) multivariable regression modeling, and (4) regression family subtype (logistic, linear, or Cox proportional hazards). Classification was based on explicit descriptions within the methods

sections of each publication. The objective was not to evaluate methodological quality but to assess structural convergence in analytical approach.

3.2.3 Results

Workflow Structure

The reconstructed workflow followed a sequential pipeline consisting of hypothesis development and interpretation, cohort specification, database query development, statistical analysis, and writing (Figure 3.2.1). Although this structure is conceptually linear, the practical execution of the workflow involved frequent feedback loops between stages. Iteration was most pronounced between cohort specification and query development, where refinements to inclusion and exclusion criteria repeatedly required modification of SQL logic and re-execution of downstream processing steps. These feedback cycles created a tightly coupled early-stage loop that dominated the overall workflow dynamics.

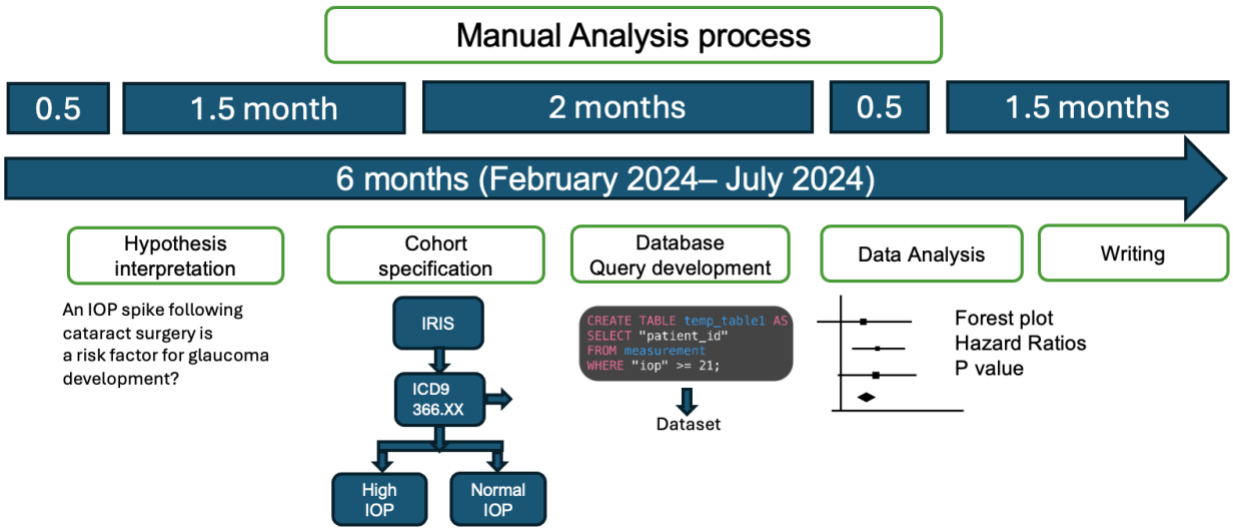


Figure 3.2.1. Manual clinical research workflow and timeline. A six-month manual pipeline for hypothesis-driven registry analysis, showing sequential stages of hypothesis development and interpretation, cohort specification, database query development, statistical analysis, and manuscript writing through initial submission.

Distribution of Time Across Workflow Stages

Time reconstruction of the manual workflow revealed an uneven distribution of effort across research stages (Figure 3.2.1). Hypothesis refinement accounted for approximately 8.3% of total project time and involved translating a clinician's natural language question into precise operational definitions, such as specifying thresholds for IOP spikes, defining postoperative time windows, and establishing criteria for glaucoma development.

Cohort specification required approximately 25% of total effort. This stage involved translating the refined hypothesis into database-aware inclusion and exclusion criteria using structured coding systems such as ICD and CPT. Decisions included defining complex eligibility rules, determining eye-level versus

patient-level analysis, handling bilateral measurements, and selecting appropriate surgical episodes. This process required repeated consultation with clinical collaborators and careful alignment with available database variables.

Database query development represented the single largest time investment at approximately 33% of total effort. This phase extended beyond simple data extraction and included extensive data cleaning, handling missing values, table merging, and iterative query refinement. Long query execution times, schema inconsistencies across dataset versions, and the availability of updated data releases required reimplementation of extraction pipelines. Communication with clinical collaborators often triggered revisions that necessitated rebuilding queries and regenerating datasets.

In contrast, statistical analysis accounted for only 8.3% of total time. Established statistical packages and standard analytical methods were used, and no new statistical techniques were developed. Reporting and writing activities represented the remaining 25%, although these tasks include wait time for collaborator review and edits.

Overall, early-stage activities, cohort specification and database query development, comprised roughly two-thirds of total effort. Iterative cycles frequently required revisiting these stages when cohort definitions evolved or when extracted data revealed structural inconsistencies. Each revision amplified the effective cost of early workflow steps by requiring repeated data extraction and preprocessing.

Reusability of Workflow Components

Reusability analysis demonstrated substantial variation across workflow components. Statistical analysis scripts were generally highly reusable because they were implemented using modular and parameterized code structures that could be adapted to related studies with minimal modification. These scripts included reusable procedures for processing common demographic variables, handling numerical, binary, and categorical data, and implementing widely used analytical methods such as survival modeling and descriptive statistical reporting. In many cases, reuse required limited adjustments, such as substituting variable names or changing reference groups.

In contrast, SQL queries used for cohort construction exhibited predominantly low to moderate reusability. Many queries were tightly coupled to study-specific inclusion and exclusion criteria, diagnosis codes, temporal filtering rules, and dataset idiosyncrasies. As a result, applying them in new contexts often required partial or substantial rewriting.

These findings suggest that while statistical procedures often generalize across studies, cohort formation tend to remain study specific. As a result, components of manual research pipelines vary in their degree of reusability, with substantial effort concentrated in logic tailored to individual study designs. Much of this work remains embedded in context-dependent implementation rather than abstracted into reusable structures. This pattern underscores the potential value of automation frameworks that emphasize modular design, reuse, and systematic translation from clinical intent to executable analysis.

Cross-Study Convergence in Statistical Methodology

Across the eight reviewed studies, statistical methodology demonstrated marked structural homogeneity. All studies (8/8, 100%) included descriptive reporting of baseline demographic and clinical characteristics. Group-stratified reporting and/or comparative statistical analyses were also universally reported (8/8, 100%). Importantly, inferential analysis in every study had some form of multivariable regression modeling (8/8, 100%). Variation across studies was limited to the specific regression family selected. Cox proportional hazards models were used in 62.5% of studies (5/8), linear regression in 50% (4/8), and logistic regression in 25% (2/8). These findings indicate that statistical analysis in this domain is confined to a standardized methodological framework consisting of descriptive reporting and multivariable regression techniques. Differences across studies arise primarily from cohort definitions, exposure specifications, and temporal filtering strategies rather than from innovation at the level of statistical modeling.

3.2.4 Discussion & Conclusion

The workflow analysis identifies dominant bottlenecks in manual clinical research pipelines that account for more than half of total project time: translating natural language clinical hypotheses into schema-aware cohort specifications and constructing SQL queries that correctly implement complex inclusion, exclusion, and temporal logic. Together, these early-stage activities represent the primary concentration of effort.

Statistical analysis stages, by contrast, rely on established and reusable methodological frameworks implemented through standardized software packages, and therefore exhibit comparatively less variability across studies. Cohort specification and database querying, however, remain highly study-specific, requiring detailed schema knowledge, iterative refinement, and precise temporal filtering. Automation efforts should therefore prioritize systems that support schema-aware translation of clinical intent and the generation of validated database operations.

Although this study examines a single real-world workflow, the structural patterns observed align with broader challenges in large observational research. Tasks within cohort definition and query construction often follow recurring procedural structures, yet they are rarely abstracted into reusable components. As a result, substantial effort remains embedded in context-specific implementation rather than generalized infrastructure. These findings provide an empirical foundation for the development of tools that reduce manual burden while preserving methodological transparency.

Importantly, the case study suggests that variability across studies is concentrated upstream in cohort definition and data preparation rather than in downstream statistical modeling. Statistical analysis in this domain largely operates within a stable methodological framework centered on descriptive reporting and multivariable regression techniques. In contrast, cohort construction requires extensive study-specific logic, schema-aware filtering, and repeated query redevelopment.

From a systems design perspective, this asymmetry has important implications. A scalable framework for registry-based research should prioritize flexibility in dataset selection, cohort specification, and query generation, where variability and manual effort are greatest, while maintaining a relatively standardized analytical layer based on established statistical methods. Modularizing the workflow in this manner,

flexible upstream cohort preparation combined with a stable downstream statistical framework, enables reduction of repetitive manual effort without altering established analytical rigor.

This design principle directly motivates the development of LLM-assisted systems that prioritize flexibility in schema-aware cohort preparation and automated query generation, while maintaining a stable and standardized statistical methodology, as discussed in the next chapter.

3.3 Chapter Summary

This chapter examined manual clinical research workflows in order to identify which components of hypothesis-driven research are most amenable to automation. Using a large IRIS Registry study of postoperative intraocular pressure elevation and glaucoma risk as a case study, the chapter first demonstrated a representative real-world clinical research analysis conducted through traditional manual methods. It then extended this work through a process analysis of the underlying workflow, highlighting repeated tasks, programming burden, bottlenecks in scalability, and other inefficiencies that limit reproducibility and speed. By making these constraints explicit, this chapter provides the practical motivation and design rationale for LATCH and lays the foundation for the framework introduced in the following chapter.

Chapter 4. The LATCH Framework: Development and Evaluation

The growth of electronic health records (EHRs) and population-scale health datasets has created unprecedented opportunities for observational discovery and clinical hypothesis testing. Yet, converting clinical questions into analyses remains slow and resource intensive[1–3], requiring domain expertise, careful cohort construction, and substantial statistical programming. These barriers limit reproducibility and impede independent verification, as analytic pipelines are rarely released in fully executable form and methodological descriptions often omit critical implementation details [4–7]. Recent advances in LLMs suggest a path toward natural language interfaces for data analysis, but naively applying LLMs to clinical research introduces serious risks, including hallucinated variables, inconsistent execution, and privacy concerns[20–22] when patient-level data are involved.

This chapter presents LATCH, an LLM-assisted framework designed to translate natural language clinical hypotheses into verifiable statistical analyses while preserving reproducibility, transparency, and privacy. LATCH is built on a hybrid architecture that separates semantic reasoning from deterministic execution: LLMs are used only where semantic reasoning is required (e.g., understanding clinical language and grounding concepts to schema, planning cohort inclusion exclusion logics), while all statistical computation, and reporting are executed through fixed pipelines in a local environment. This design enables flexible interaction through natural language without sacrificing the methodological rigor required for biomedical research. In addition, LATCH produces a comprehensive Analytic Report that records each intermediate artifact, from the parsed hypothesis, used variables, executed database code, analysis code, and final analysis outputs, creating an auditable, end-to-end record of the full analytic workflow.

The remainder of this chapter discusses LATCH from three complementary perspectives. Section 4.1 describes the system architecture, modular components, and safeguards that enforce schema grounding and mitigate failure modes during LLM-assisted steps. Section 4.2 validates analytical fidelity by reproducing a curated set of published NHANES-based diabetes studies, quantifying concordance with reported results and using reproduction failures to identify sources of ambiguity in the literature. Section 4.3 then characterizes operational limits through controlled stress testing, such as behavior upon missing or contradictory specifications, semantic variation between clinical terminology and dataset schemas, logical complexity in cohort construction and feature engineering, and variability across alternative LLM backends. Together, these studies establish both the practical utility of LATCH for accelerating clinical hypothesis testing and the boundary conditions required for reliable deployment in scientific workflows. Portions of this chapter are adapted from the author's preprint[22].

4.1 System Architecture: LLM-assisted Semantic Layer with Deterministic Statistical Engine

4.1.1 Introduction

LLMs have recently demonstrated strong capabilities in natural language understanding and code generation, raising the possibility of automating portions of the clinical research pipeline. Yet the direct application of LLMs to patient-level data poses substantial challenges, including risks of hallucinated variables or methods, inconsistent analytic execution, and concerns related to data privacy and regulatory compliance.

To address these challenges, we developed LATCH, an LLM-assisted framework that translates free-text clinical hypotheses into verifiable statistical analyses over structured health datasets. LATCH is built on a hybrid architecture that separates semantic reasoning from deterministic execution. LLMs are used selectively for tasks that require semantic understanding of natural language and schema grounding, while all statistical computation and data handling are performed through fixed, reproducible pipelines within a secure local environment. This design enables flexible natural language interaction without sacrificing the rigor required for clinical research. In this section, we describe the system architecture of LATCH, including its modular design, safeguard mechanisms, and process-logging features. We then demonstrate how this architecture supports hypothesis testing from structured EHR data while maintaining strict privacy guarantees and transparent audit trails.

4.1.2 Methods

Overview

LATCH is an automated framework designed to streamline the analysis of tabular health data through a standardized and transparent pipeline. A central design philosophy of the system is the selective use of LLMs. The system employs LLMs exclusively for tasks requiring semantic reasoning, such as interpreting ambiguous clinical language, while utilizing deterministic, rule-based systems for statistical execution and recording the process. This hybrid approach allows LATCH to maintain the flexibility of

natural language understanding while adhering to the strict reproducibility standards required for medical research.

Architecture

LATCH comprises five specialized modules that mirror the standard scientific workflow: Planner, Variable Mapper, Data Engine, Statistics Engine and Reporter. These modules are organized into two functional layers. The first consists of three hybrid semantic modules (Planner, Variable Mapper and Data Engine), which integrate LLM-based semantic reasoning with rule-based constraint enforcement to generate executable code and logic from natural language input. The second layer comprises two fully deterministic execution modules (Statistics Engine and Reporter), which perform statistical computation, aggregation and reporting using fixed, reproducible pipelines.

A central Python orchestration layer coordinates this workflow, integrating heterogeneous software components, such as LLM APIs, PostgreSQL[119], and R, via libraries such as SQLAlchemy[120] and rpy2. Crucially, this architecture enforces data privacy by design. The LLM is used only to generate code and logic, and it never accesses the patient database directly. Data processing occurs locally within the secure environment, ensuring that patient-level information is never exposed to external API calls.

All components within the LLM-assisted semantic layer (Planner, Variable Mapper, and Data Engine) were driven by a structured prompt template that combined step-by-step task instructions with a one-shot example[28] demonstrating the expected input-output format. This standardized prompting strategy improved output consistency, reduced formatting and logical errors, and enforced structured, machine-readable responses.

Planner

The Planner module uses a large language model (Google Gemini 2.5 Flash; temperature = 0) in two sequential processing stages. In the first stage, free-text study hypotheses are standardized into a structured text with a field-based analytical specification. This process extracts key analytical components, including the selected dataset, study period, analysis method, and method-specific variables (for example, predictors, covariates, and outcomes), as required by the analysis type. During this step, the system's safeguard identifies missing or inconsistent study specifications, flags unsupported analytical methods, and prompts users for clarification when necessary. Subsequently, the structured text is converted into a machine-readable JSON study plan. This dual-stage design improves JSON generation reliability and ensures that the study plan is aligned with the requirements of the selected analytical model.

Variable Mapper

The Variable Mapper module connects clinical concepts in the study plan to database variables. To prevent hallucinations, this module uses a hybrid retrieval process rather than direct LLM generation. First, keywords are extracted from the JSON study plan in a rule-based manner. For each keyword, a semantic search model (all-MiniLM-L6-v2[121]) retrieves candidate variables from the schema summary based on cosine similarity. Because the schema summary contains a large number of variables across the

full dataset, this distance-based filtering step is used to identify the top 40 candidate variables and to ensure compatibility with LLM context window constraints. Second, an LLM performs a classification to select the most contextually appropriate variable from this set of candidates. Each candidate variable is presented to the LLM as a filtered subset of the schema summary table, with available metadata fields such as table name, variable name, example values, and descriptions, allowing the model to compare candidates and select the best-aligned variable.

This process is performed for each keyword in the study plan and produces a validated mapping between clinical concepts (e.g., "fasting blood sugar level") in the user query and database variable (e.g., fasting_glucose_mmol_l), ensuring that downstream analyses are grounded in the available schema. These mappings are incorporated into the JSON study plan in a rule-based manner, resulting in an enriched specification in which each keyword is paired to its corresponding table and column identifiers.

Data Engine

The Data Engine module ingests the enriched JSON study plan to programmatically construct the analytic cohort. It begins with a rule-based step that aggregates mapped variables across data cycles and produces a master table. Performing this step outside the LLM ensures structural integrity, reduces context window overhead, and reserves LLM compute for higher-level reasoning, as the procedure involves only deterministic variable aggregation. Subsequently, the schema of the master table is provided to the LLM, which functions as a logic synthesizer and generates SQL queries to implement study-specific inclusion/exclusion criteria, perform feature engineering, and format variables according to the required statistical analysis method. The generated SQL query is executed locally via SQLAlchemy, and results are stored as a pandas DataFrame.

Statistics Engine

The Statistics Engine module executes the statistical analysis on the prepared pandas[62] DataFrame outputted by the SQL query. LATCH has the separation of the flexible, semantic query interpretation that involve LLM-assisted steps from rigid statistical execution. This design ensures consistency and reproducibility by eliminating analytic variability, such as inconsistent statistical packages, parameters, or significance thresholds, which would invalidate comparisons across separate analyses.

The module processes the input data through a sequential, automated analytical pipeline. On the pandas DataFrame outputted by the SQL query, feature classification occurs to label variables as numerical, binary, or categorical. If needed, reference levels for categorical predictors are automatically selected based on clinically meaningful keywords (e.g., "healthy", "none", "control", "Q1") or, when absent, the most frequent level. Statistical analyses are executed according to the analysis method specified during the Planner phase, which was selected from a predefined set of available statistical analyses. Supported analyses include logistic regression, linear regression, Cox proportional hazards, prevalence estimation, group-wise numerical comparisons, stratified logistic regression, and mediation analysis.

For unweighted analyses, the module utilizes lm for linear regression, glm (binomial family) for logistic regression, and coxph from R's survival[122] package for time-to-event modeling. Group-based comparisons of continuous variables are conducted using Welch's t-test for two-group comparisons and

one-way Analysis of Variance (ANOVA) for multi-group scenarios. Categorical associations are evaluated using Pearson's chi-square test. Mediation analyses are performed using the mediate package. All confidence intervals are calculated at the 95% level, and effect estimates are reported as odds ratios (OR) or hazard ratios (HR) where applicable. For complex survey designs (e.g., NHANES), the module utilizes the R survey[123] package to account for multi-stage sampling. Regression analyses are executed via svyglm and svycoxph. Weighted prevalence estimates and group-specific summary statistics are computed using svymean and svyby. Between-group differences are evaluated using design-based hypothesis tests, including svyttest for two-group comparisons and svychisq for categorical group comparisons. Survey weighting is enforced through fixed, deterministic templates applied as a rule-based layer on top of the standard query generation pipeline, conforming to NHANES analytic guidelines and automatically filtering participants without valid sampling weights while ensuring inclusion of all required survey design variables prior to statistical execution.

In case of missing data, for survey-weighted analyses, the engine defaults to complete-case analysis to maintain the integrity of the survey design structure, which can be compromised under standard imputation. For unweighted analyses, complete-case analysis is the default, with optional support for Multiple Imputation via Chained Equations (MICE)[124] or simple imputation (mean/mode) for exploratory sensitivity analyses.

Reporter

The final module generates a comprehensive Analytic Report that serves as an immutable audit trail of the full analytical workflow. This component operates as a rule-based logging system that records all intermediate and final pipeline artifacts in a structured comma-separated values (CSV) format. Each column in the report corresponds to a predefined metadata field, and all inputs and outputs are captured without modification. The report aggregates structured outputs from each stage of the pipeline, including the parsed user hypothesis, cohort definition parameters, executed SQL queries, selected schema variables, and final statistical results such as model coefficients and summary tables. Each field is written to predefined columns using fixed formatting rules, ensuring that all inputs and outputs remain directly traceable, verifiable, and reproducible.

Safeguard System

A safeguard mechanism is integrated across the LLM-assisted components of LATCH, specifically the Planner, Variable Mapper, and Data Engine, to enhance operational robustness.. The safeguards are applied at three LLM-assisted stages of the workflow, as described below:

Safeguard 1. At the Planner step, this performs user input vetting, evaluating the user's free-text request against high-level study design requirements (e.g., compatibility between analysis method and outcome type) and verifying dataset and study-year availability. Requests that are ambiguous, incomplete, or unsupported trigger a clarification prompt with specific guidance before execution proceeds. Safeguard 2. During the Variable Mapping step, this applies a rule-based filter that first identifies candidate variables that are consistently defined across the specified study period, mitigating errors due to variable drift across data cycles. When multiple consistent candidates exist, an LLM selects the most contextually appropriate variable based on the full study specification; when no single consistent variable is available,

year-specific mappings are applied and logged for transparency. Safeguard 3. In the Data Engine module, this performs iterative SQL Self-Correction in the Data Engine module. If a generated SQL query fails execution, the database error message and schema context are fed back to the LLM to refine the query. This automated correction loop iterates for up to three attempts, resolving syntax or logic errors before terminating.

4.1.3 Results

LATCH: An LLM-assisted framework for verifiable clinical hypothesis testing from EHR data

LATCH is a framework that translates natural language clinical questions into reproducible statistical analyses with structured EHR data. Its architecture strictly isolates LLM-integrated steps from all patient-level operations, ensuring that no patient-level data are exposed to LLMs during any stage of analysis. LATCH is designed to operate on any standardized structured health dataset for which a data dictionary or schema is available. This requires a one-time setup that allows the model to understand the dataset's variables and their meanings, schema, without accessing patient-level data. This preparation has two components (Figure 2.1.1). First, the Code-to-Text Translation step translates coded variable names into natural language descriptions. This not only allows the LLM to accurately infer the meaning of each variable without prior knowledge of the coding system but also ensures that generated analytic code is interpretable during human review. Second, the Schema Summary Table is constructed to provide an overview of available variables and their attributes. During routine operation, only this schema-level metadata and not patient-level data is shared with the LLM. The resulting schema summary provides the foundation for subsequent schema-grounding steps, enabling LATCH to map natural language concepts to database variables.

LATCH comprises five specialized modules: Planner, Variable Mapper, Data Engine, Statistics Engine, and Reporter (Figure 4.1.1a, 4.1.2). These modules are organized into an LLM-assisted semantic layer, which supports code and logic generation from natural language input, and a deterministic execution layer which performs statistical computation and reporting. In this workflow, the Planner converts the clinical question into a structured analysis specification, the Variable Mapper matches this specification to dataset variables using the schema summary, and the Data Engine generates executable database queries to extract the study cohort. The extracted dataset is processed by the Statistics Engine, and the Reporter produces a comprehensive Analytic Report documenting natural language query, schema mappings, executed code, cohort characteristics, and results (Figure 4.1.3).

To mitigate errors, LATCH incorporates three safeguards for each of the LLM-integrated steps. First, during the Planner stage, it checks whether the user's question is missing critical study details or contains internal contradictions. Second, during the Variable Mapping stage, it verifies that selected variables are consistent across survey years and flags any implausible mappings. Third, for the Data Engine, if the generated SQL fails to execute, LATCH automatically repairs and retries up to three times. All analyses are conducted with safeguards enabled, and their impact on error mitigation is evaluated separately, using targeted perturbations of inputs to isolate the effects of each safeguard. Figure 4.1.1b provides an overview of LATCH applications, which are discussed in depth throughout the remainder of Chapter 4 and in Chapter 5

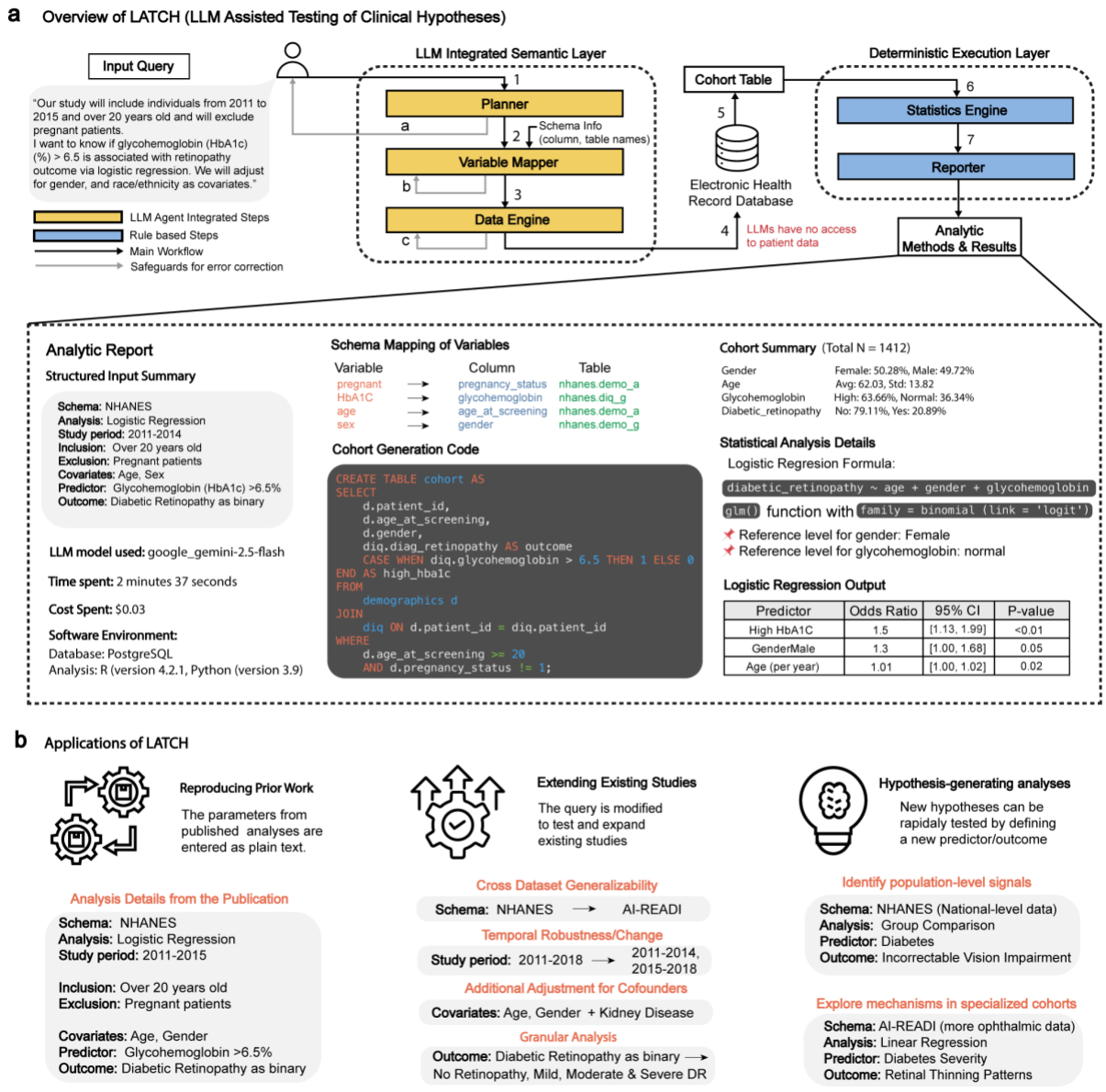


Figure 4.1.1. Overview of LATCH framework and its applications. a, Architecture and workflow of LATCH. (1) A user's natural language query is processed by the Planner module to generate a structured analysis specification. (2) This specification is passed to the Variable Mapper, which incorporates dataset schema metadata to produce schema-aligned variable mappings. (3) The mapped query is translated into executable database code by the Data Engine to define the study cohort and extract required variables. (4) All cohort extraction and data processing are performed locally, ensuring that patient-level data are never exposed to LLMs. (5) The extracted dataset is (6) analyzed by a deterministic Statistics Engine. (7) Outputs, including inputs, mappings, executed code, cohort summaries, and statistical results, are recorded by the Reporter module to generate an immutable Analytic Report. Safeguards are applied across LLM-integrated steps (a-c) to detect and mitigate potential failure modes. **b,** Key potential applications of LATCH include reproducing published analyses from text-based study specifications (left), extending existing studies (middle), and conducting hypothesis-generating analyses by testing previously unexplored text-based study specifications that define and evaluate new predictors and outcomes (right). Reproduced from ref. 22 under a CC BY 4.0 license.

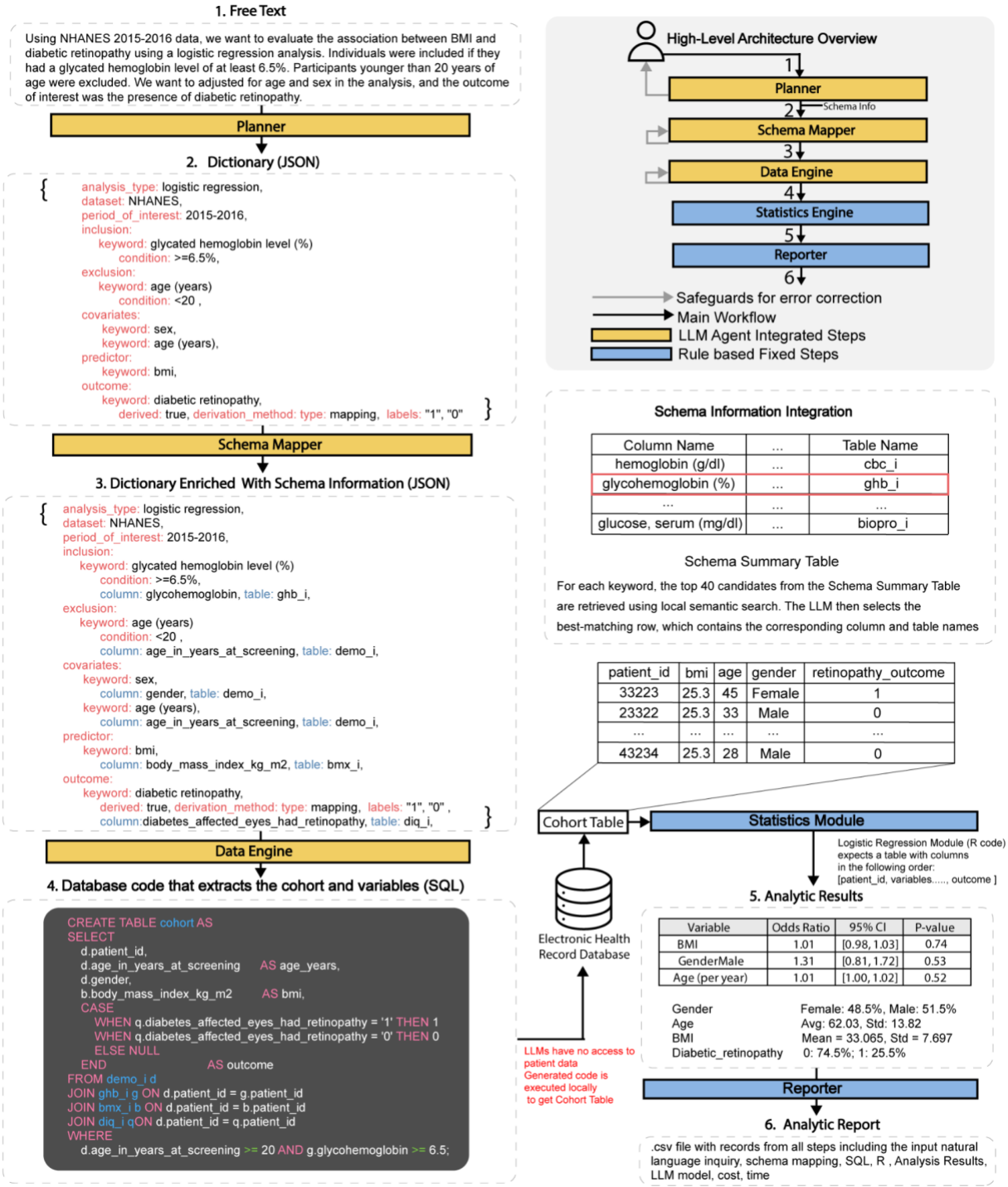


Figure 4.1.2. | Stepwise LATCH analytic pipeline. (1) A natural language research question is converted into (2) a structured JSON encoding cohort definitions and analytic parameters. This specification is (3) enriched by mapping concepts to dataset-specific schema elements and used to generate (4) executable SQL that extracts the cohort. The resulting table is processed by statistical modules to produce (5) summaries and regression outputs. All workflow stages are documented in (6) an Analytic Report. All code (SQL and R) is executed locally, and LLMs do not access patient-level data. Reproduced from ref. 22 under a CC BY 4.0 license.

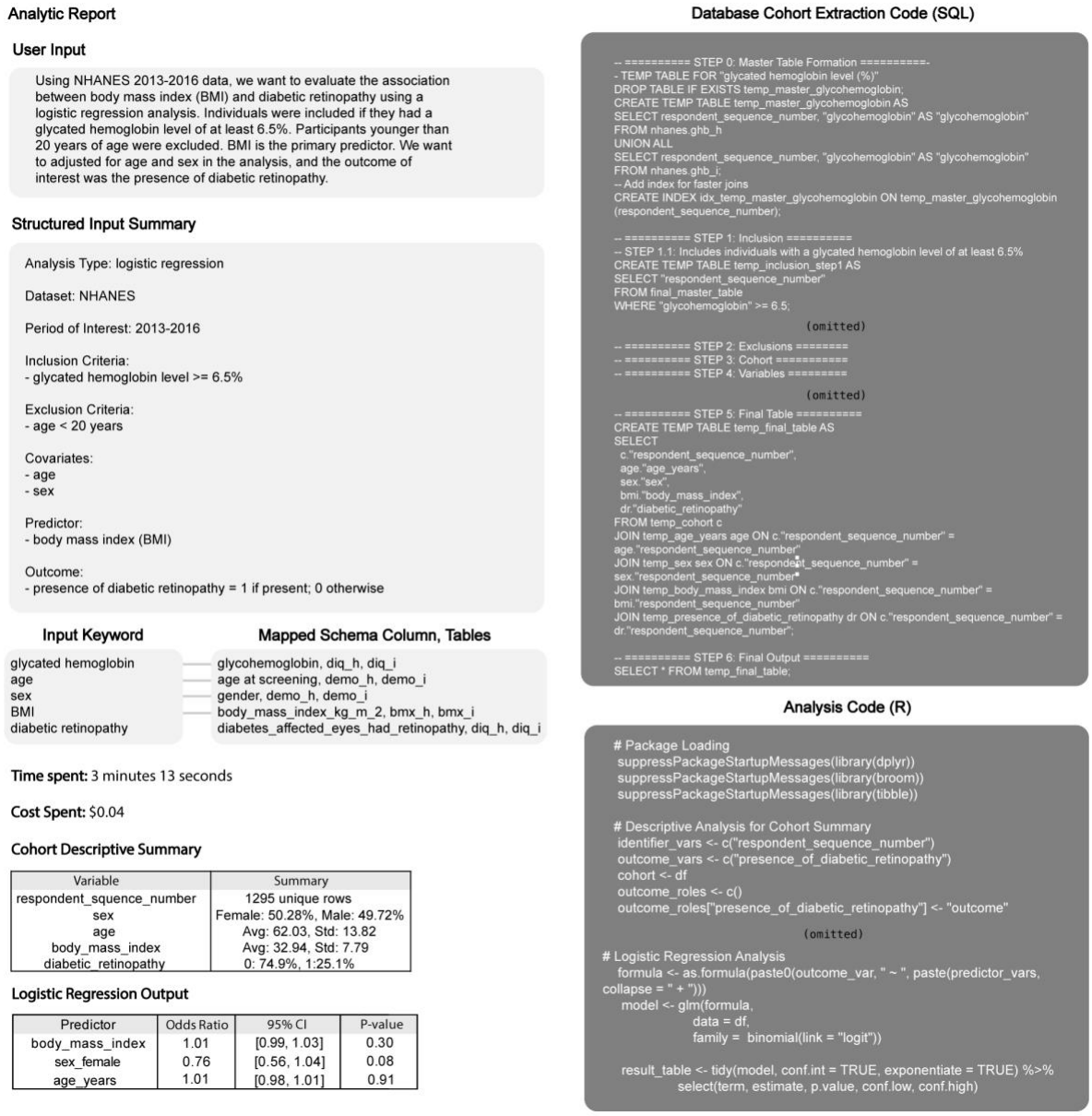


Figure 4.1.3. Example analytic report generated by LATCH. Illustrative analytic receipt documenting an end-to-end analysis generated from a natural language research query. The receipt includes the original user input, a structured study design summary specifying analysis type, dataset, inclusion and exclusion criteria, covariates, predictor, and outcome, and the corresponding schema mappings matching user-specified concepts to database tables and columns. Generated cohort extraction code (SQL) is shown alongside downstream statistical analysis code (R), enabling transparent reconstruction of cohort definition and modeling steps. Execution time, API cost, cohort characteristics, and regression outputs, are recorded to support auditability, reproducibility, and tracking of each analysis. Reproduced from ref. 22 under a CC BY 4.0 license.

4.1.4 Discussion & Conclusion

The LATCH framework demonstrates how LLMs can be integrated into clinical data analysis pipelines in a manner that preserves reproducibility, transparency, and privacy. By isolating LLMs within a semantic

interpretation layer and delegating all statistical execution to deterministic engines, LATCH addresses a central tension in AI-assisted research: balancing flexibility in natural language interaction with the stringent methodological standards required for biomedical science.

A key strength of LATCH lies in its modular architecture. The separation into Planner, Variable Mapper, Data Engine, Statistics Engine, and Reporter mirrors the conventional scientific workflow and enables independent validation of each stage. The incorporation of safeguard mechanisms at LLM-assisted step further improve robustness by various mechanisms, such as detecting ambiguous specifications, preventing schema mismatches, and automatically correcting executable errors. Together, these features produce an auditable analytic trail in which every transformation, from natural language query to final statistical output, is documented and reproducible.

Several limitations should be acknowledged. First, LATCH currently supports a predefined set of statistical methods, which, while covering many common epidemiological analyses, does not encompass the full spectrum of advanced modeling techniques. Second, the accuracy of variable mapping depends on the quality and completeness of the dataset schema summary. Inconsistent or poorly documented metadata may limit performance. Third, although safeguard mechanisms reduce errors, they cannot eliminate all ambiguities inherent in natural language specifications, and expert oversight remains important for high-stakes analyses.

In conclusion, LATCH provides a practical blueprint for integrating LLMs into clinical research workflows without compromising reproducibility or privacy. By combining semantic flexibility with deterministic execution and comprehensive audit logging, the framework enables transparent and scalable hypothesis testing from structured health data.

4.2 Validation through Reproduction of Published Studies

4.2.1 Introduction

Reproducibility is a cornerstone of scientific credibility, yet large-scale observational studies built on complex public health datasets remain challenging to replicate in practice. The NHANES is one of the most widely used epidemiological resources for studying diabetes and its complications, supporting hundreds of publications that inform clinical understanding and public health policy. However, reproducing NHANES-based analyses typically requires specialized expertise in survey design, data harmonization across cycles, and statistical modeling, as well as careful interpretation of often underspecified methodological descriptions in published work. These barriers limit independent verification of findings and slow the translation of research insights into robust, reusable knowledge.

Recent advances in LLMs offer the possibility of automating complex analytic workflows through natural language instructions. LATCH is designed to operationalize this capability by translating structured prompts into standardized cohort construction, data processing, and statistical analyses on large biomedical datasets. While automated systems promise to accelerate research and democratize access to advanced analytics, their reliability must be rigorously validated against established scientific results.

In this section, we evaluate LATCH by reproducing a curated set of published NHANES-based diabetes studies spanning multiple clinical domains and analytic paradigms. By systematically comparing LATCH-generated outputs with original published findings using predefined concordance criteria, we aim to establish a quantitative benchmark for analytical fidelity, identify sources of reproducibility challenges in the literature, and assess the consistency of AI-assisted analytic pipelines. This evaluation provides insight into both the feasibility of automated reproduction and the broader role of LLM-driven systems in strengthening scientific reproducibility**.**

4.2.2 Methods

Literature Search and Curation of Reproduction Studies

To evaluate the reproducibility of NHANES-based diabetes research, we curated a representative cohort of 20 studies (Figure 4.2.1). PubMed was searched for articles published between 2010 and 2025 using combinations of the following terms: ("National Health and Nutrition Examination Survey" OR "NHANES") AND "diabetes", together with outcome-specific keywords for diabetic retinopathy ("diabetic retinopathy"), diabetic kidney disease ("diabetic kidney disease"), all-cause mortality("all-cause mortality"), and mental or cognitive health ("depression" OR "cognitive" OR "mental" OR "cognition"). Search terms restricted to title and abstract fields. Inclusion was limited to studies with full-text availability to ensure the presence of sufficient methodological detail for reproduction.

Because the objective was to assess reproduction feasibility rather than provide a comprehensive meta-analysis, we employed a ranked-sampling approach. The retrieved articles per topic were ranked by Google Scholar citation count in descending order. We screened these results sequentially and applied the following exclusion criteria: use of non-standard or restricted NHANES data or external data sources; populations not restricted to individuals with diabetes; redundancy in exposure-outcome relationships already represented by a higher-ranked study; focus on outcomes outside the four domains of interest; methodological descriptions insufficient for exact reproduction, including ambiguous definitions of inclusion/exclusion criteria, exposures, outcomes, or covariates; involved data structures unsuitable for standardized multi-cycle analysis, such as highly granular dietary or prescription drug records; and statistical methods unsupported by the LATCH framework.

Following this ranked screening procedure, the first five eligible studies were selected for each of the diabetic retinopathy, diabetic kidney disease, and mental or cognitive health outcome categories. To demonstrate time-to-event analyses (e.g., Cox regression), three all-cause mortality studies were included. In addition, two supplementary diabetic retinopathy studies were incorporated because they contained overlapping risk factor definitions shared between the NHANES and AI-READI datasets, enabling planned cross-dataset validation experiments. In total, 20 studies were selected for reproduction, and the distribution of medical topics, statistical analysis types, publishers, and journal impact factors is summarized in Figure 4.2.1.

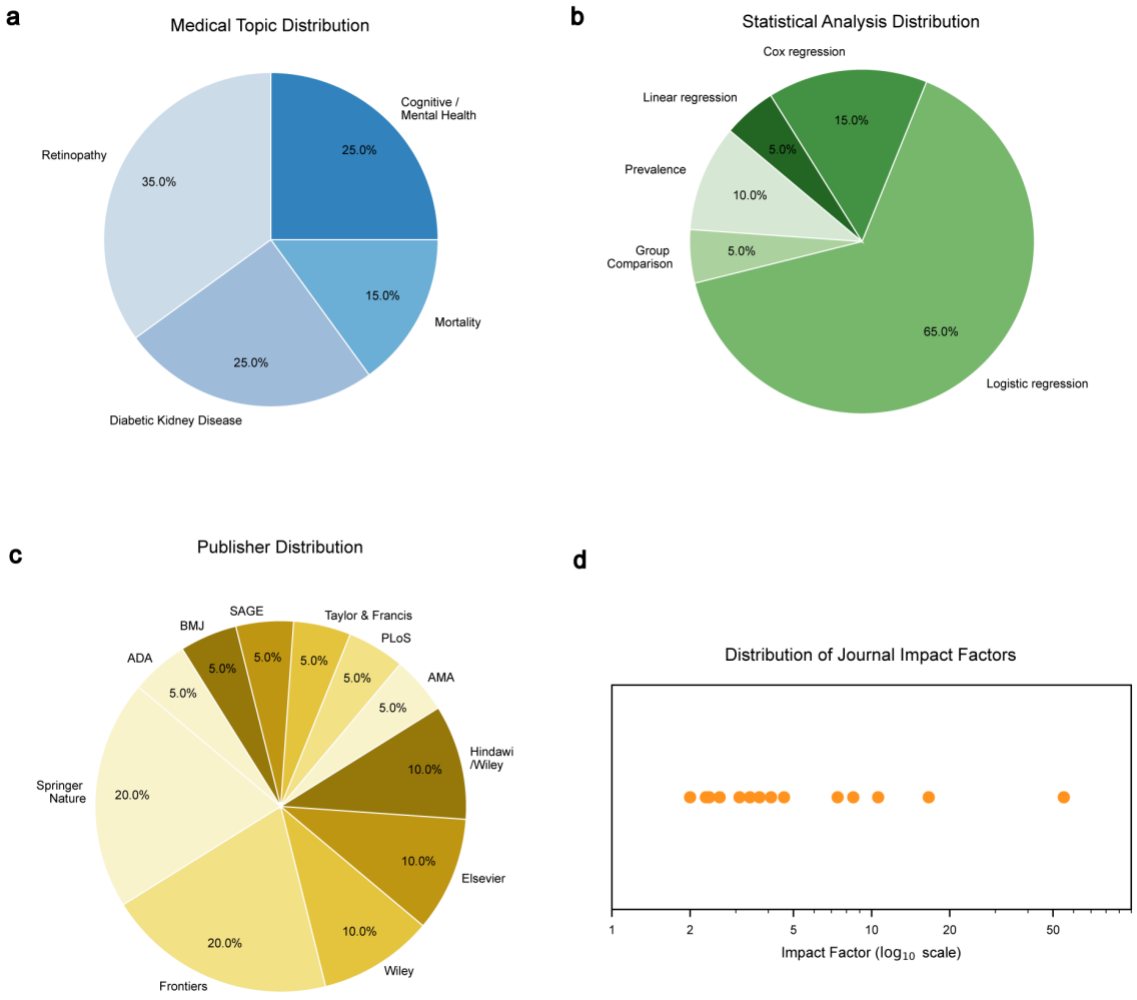


Figure 4.2.1. Characteristics of the reproduction studies included in the evaluation. a, Medical topic distribution of the 20 selected replication studies, spanning diabetic retinopathy, diabetic kidney disease, cognitive or mental health outcomes, and all-cause mortality. **b,** Statistical analysis distribution of the analyses LATCH reproduced. **c,** Publisher distribution of the reproduction studies, including a range of biomedical journals and publishers. **d,** Distribution of journal impact factors for the reproduction studies, shown on the raw scale with logarithmic spacing. Reproduced from ref. 22 under a CC BY 4.0 license.

Running Reproduction Analysis on LATCH

The primary objective of this reproduction study was to assess LATCH's ability to reproduce previously published results. We began by systematically decomposing the methodologies of the original publications into a structured framework comprising cohort definitions, inclusion/exclusion criteria, predictor and outcome variables (if applicable), and variable transformations. For all derived variables, we explicitly defined the computational logic and underlying raw data components within the prompt to ensure the LATCH system accurately reconstructed the original study's metrics. These parameters were synthesized into natural language prompts to guide the LATCH system's automated execution, ensuring that each analysis remained strictly aligned with the original study periods and datasets.

While the original publications typically reported results across three levels of complexity, univariate, demographic-adjusted, and fully adjusted models, we focused our primary reproduction on the demographic-adjusted models to ensure a high degree of analytical comparability. By prioritizing demographic models over "fully adjusted" specifications, which often involve heterogeneous covariate definitions and under-reported missing data handling protocols, we minimized the risk of misattributing discrepancies to irreproducibility when they may have stemmed from unstated modeling choices. We standardized these models using a consistent set of variables, specifically “Gender,” “Age at screening,” and “Race/Hispanic Origin,” and ensured all laboratory variables included units. This standardized approach provided a uniform baseline for validation. For the two studies [125–127] where a demographic-only model was unavailable, we utilized the most conservative (least-adjusted) specification reported to ensure a fair assessment of the original findings.

For studies utilizing complex survey designs, we specified survey weights within the prompt as described in the original publications. In cases of methodological ambiguity, we prioritized a conservative default (e.g., examination weights if exam values were used as a variable) to ensure the analytic sample size remained consistent with the source material and to prevent artificial sample loss. The reproduction was performed by inputting these individual prompts, one per analysis, directly into the LATCH system; the subsequent data retrieval, processing, statistical analysis, and results generation were automated.

Concordance Evaluation

Concordance between results generated by LATCH and those reported in the original publications was evaluated using multiple complementary metrics designed to capture agreement in direction, statistical significance, effect magnitude, and cohort construction.

Primary concordance measures were applied primarily to logistic regression and Cox proportional hazards models, which constituted the majority of reproduced analyses and provided standardized relative effect estimates suitable for direct comparison. These measures included assessment of effect direction, determining whether both analyses indicated increased or decreased risk; statistical significance concordance, evaluating whether both results were significant ($p < 0.05$) or both were non-significant; and overlap of 95% confidence intervals when available, focusing on the most extreme exposure category (for example, the highest quartile or tertile). Additional descriptive measures were used to support interpretation, including comparison of analytic sample sizes between LATCH and the published studies to assess alignment in cohort construction, and calculation of relative effect size ratios (published estimate divided by the LATCH estimate) when effect measures were directly comparable[128].

When discordance was observed, a root-cause analysis was conducted by manually reviewing both the constructed prompts and the original publications to identify potential sources of discrepancy, such as ambiguous cohort definitions, unclear variable transformations, or insufficiently specified data processing procedures.

Because outcome reporting formats varied substantially across studies, concordance criteria were adapted accordingly. For prevalence or percentage-based outcomes lacking confidence intervals, agreement was assessed using absolute differences, with values considered concordant if the difference was within 1 percentage point. For group comparisons reported only as means and standard deviations, concordance

focused on agreement in direction of between-group difference and confidence interval overlap when available. For linear regression analyses reporting beta coefficients with confidence intervals and p-values, concordance was evaluated based on effect direction, confidence interval overlap, and statistical significance, while relative effect size ratios were applied only when effect scales were directly comparable.

After applying the concordance criteria specific to each analysis type, a replication outcome was assigned for each reproduced analysis. An analysis was classified as fully concordant if all applicable replication assessment criteria were satisfied. If one or more criteria were met but at least one criterion failed, the result was classified as partially concordant. Analyses failing all primary concordance criteria were classified as non-concordant.

Repetition Study

To assess output consistency and sources of variability, we executed each of the reproduction analyses from 20 studies independently three times using LATCH under identical input conditions. Reproduction outcomes were classified as consistent when all runs produced equivalent cohort definitions and statistical results, and as variable when discrepancies were observed across runs. For variable cases, discrepancies were manually reviewed and categorized into three sources: Valid interpretability that reflects reasonable alternative implementations of analytic details not explicitly specified in the original publication; Underspecified methods where insufficient methodological detail precluded a unique reconstruction; Misimplementation which corresponds to incorrect application of clearly defined analytic rules.

4.2.3 Results

Application of LATCH for Reproducing Prior Findings

To evaluate LATCH's utility, we applied the framework to reproduce the findings of published NHANES-based diabetes studies. Twenty publications were systematically selected based on methodological clarity, citation impact, and relevance to diabetes-related health outcomes. These studies span diverse clinical domains, including diabetic retinopathy[129–135], diabetic kidney disease[125,136–139], mental/cognitive health[126,127,140–142], and all-cause mortality[143–145], and represent a range of statistical methodologies, including logistic and linear regression, Cox proportional hazards models, and descriptive statistics (Figure 4.2.1, Figure 4.2,2, Table 4.2.1, Table 4.2.2). For each study, key methodological details from the manuscripts, including inclusion criteria, exclusion criteria, analysis methods, and used variables, were extracted and rephrased into natural language prompt inputs for LATCH.

Side-by-side comparisons of published results and LATCH-generated outputs are shown in Figure 4.2.2. The concordance between LATCH and published findings was subsequently evaluated using effect direction, statistical significance, and effect size agreement, with summary concordance metrics[128] shown in Table 4.2.1 and Table 4.2.2.

In logistic regression analyses (Figure 4.2.2a), LATCH reproduced both the direction and magnitude of many published odds ratios (ORs) for diabetic retinopathy. For example, the association between monocyte-to-lymphocyte ratio and diabetic retinopathy was closely matched (published OR = 1.59 (n =

367), 95% confidence interval (CI): 1.22−2.07; LATCH OR = 1.59 (n = 367), 95% CI: 1.22−2.09). The LATCH result reported a more precise P value that was rounded in the original publication (published P $\leq$ 0.001; LATCH P $\leq 7.94 \times 10^{-4}$). LATCH also reproduced protective associations of serum albumin and hemoglobin with diabetic retinopathy, as well as reported risk gradients for diabetic kidney disease and depression across various risk factors (Figure 4.2.2a). In Cox proportional hazards models (Figure 4.2.2b), LATCH identified elevated all-cause mortality risk among patients with diabetes associated with higher uric acid and red blood cell distribution width. Beyond regression, prevalence estimates (Figure 4.2.2c), descriptive comparisons of numerical variables (Figure 4.2.2d), and linear regression analyses (Figure 4.2.2e) were reproduced. Reconstructed cohort sizes matched those reported in the original studies, with most differing by less than 10% and several matching exactly (Table 4.2.1). End-to-end reproductions were completed efficiently, with runtimes ranging from 3 to 15 minutes.

Table 4.2.1 presents detailed concordance results for logistic and Cox regression analyses only, as the other analysis types required different concordance criteria. Table 4.2.2 summarizes overall study-level concordance across all 20 reproduced studies. Studies were classified as fully concordant when all metrics listed in the Reproduction Assessment Bases column of Table 4.2.2, applicable to the given analysis type, were concordant. Studies were classified as partially concordant when only a subset of these metrics were concordant, and as non-concordant when none of the relevant metrics were concordant. As summarized in Table 4.2.2, 16 of the 20 reproduced studies were fully concordant and 4 were partially concordant. No study was classified as fully non-concordant. Among the four partially concordant studies, two discrepancies were most plausibly explained by ambiguity in the published inclusion criteria, whereas the other two appeared to reflect uncertainty in selecting among multiple similar variables. Although these factors provide plausible explanations based on available documentation, the lack of access to original analysis code precludes definitive attribution of the underlying causes.

Importantly, these differences should not be interpreted as evidence that the original studies were irreproducible. Rather, they reflect common challenges in reproducing published analyses when methodological details are incomplete, ambiguous, or open to more than one reasonable interpretation, along with common reproducibility challenges such as undocumented preprocessing decisions, and software implementation differences. Moreover, the studies included in this evaluation were already selectively chosen because, as described in the methods section, they satisfied predefined reporting requirements. These include sufficiently clear inclusion and exclusion criteria, dataset years, variable definitions, and numerical cutoffs. As such, this evaluation likely represents a relatively favorable subset of the literature. Viewed in this context, the remaining discrepancies further underscore the value of frameworks such as LATCH, which make analytic assumptions more explicit and provide a more structured basis for transparent and reproducible study implementation.

To evaluate the stability and reproducibility of LATCH-generated analyses given the non-deterministic nature of large LLMs, we conducted repeated reproductions (n= 3) of 20 studies (Figure 4.2.3). LATCH produced consistent outputs in most cases, with 17 studies fully concordant and 3 showing variation, primarily due to reasonable but different methodological interpretation, and one case resulting from misimplementation. For these cases, Figure 4.2.2. shows the most frequent output. Together, these findings highlight the importance of human oversight to ensure system-generated analyses align with the intended analysis.

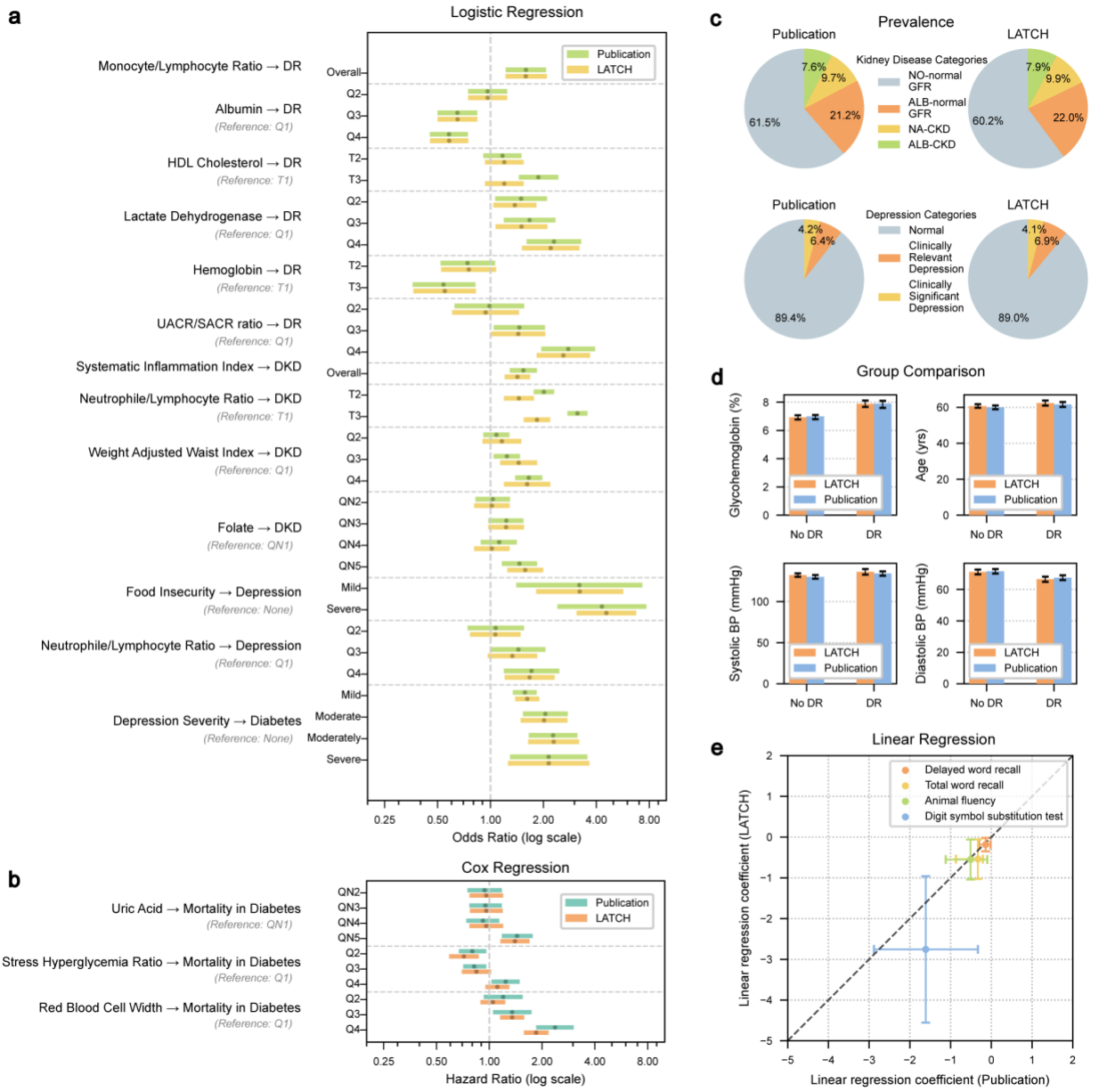


Figure 4.2.2. Reproduction of published studies. Dots indicate point estimates, with error bars representing 95% confidence intervals (CIs). **a,** Logistic regression odds ratios between published results and LATCH results. **b,** Cox regression hazard ratios between published results and LATCH results. **c,** Prevalence of different kidney disease types and depression severity in diabetes between published results. **d,** Group comparisons between patients with and without diabetic retinopathy ratios between published results and LATCH results. **e,** Linear regression coefficients for the effect of glycohemoglobin on cognitive scores as reported in published results (x axis) versus those calculated by LATCH (y axis). Abbreviations: QN, quintile; Q, quartile; T, tertile; DR, diabetic retinopathy; DKD, diabetic kidney disease; HDL, high-density lipoprotein; UACR/SACR, urinary serum albumin-to-creatinine ratio; NA-normal GFR, normal albuminuria normal glomerular filtration rate; ALB-normal GFR, albuminuria with normal glomerular filtration rate; NA-CKD, normal albuminuria chronic kidney disease defined by glomerular filtration rate; ALB-CKD, albuminuria with chronic kidney disease. Concordance evaluation results across different metrics are provided in Table 4.2.1 and Table 4.2.2. Reproduced from ref. 22 under a CC BY 4.0 license.

Analysis	Regression Type	Sample Size (Pub / LATCH)	Effect Direction Concordant	P value Concordant	CI Overlap	Relative Effect (Pub / LATCH)	Time (min)
Monocyte to Lymphocyte Ratio → Diabetic Retinopathy	Logistic	367 / 367	✓	✓	✓	1.00	3.1
Albumin → Diabetic Retinopathy	Logistic	2,964 / 2,964	✓	✓	✓	1.00	4.0
HDL Cholesterol → Diabetic Retinopathy	Logistic	1,708 / 1,380	✓	✗	✓	1.56	3.4
Lactate Dehydrogenase → Diabetic Retinopathy	Logistic	3,476 / 3,476	✓	✓	✓	1.04	3.8
Hemoglobin → Diabetic Retinopathy	Logistic	837 / 844	✓	✓	✓	0.98	1.6
UACR/SACR Ratio → Diabetic Retinopathy	Logistic	2,594 / 2,594	✓	✓	✓	1.07	5.7
Systematic Inflammation Index → Diabetic Kidney Disease	Logistic	3,937 / 3,822	✓	✓	✓	1.08	6.5
Neutrophile to Lymphocyte Ratio → Diabetic Kidney Disease	Logistic	7,153 / 7,048	✓	✓	✗	1.70	15.1
Weight Adjusted Waist Index → Diabetic Kidney Disease	Logistic	5,028 / 5,155	✓	✓	✓	1.02	8.0
Folate → Diabetic Kidney Disease	Logistic	3,461 / 3,396	✓	✓	✓	0.93	4.9
Food Insecurity → Depression in Diabetes	Logistic	1,724 / 1,688	✓	✓	✓	0.94	3.7
Neutrophile to Lymphocyte Ratio → Depression in Diabetes	Logistic	2,820 / 2,703	✓	✓	✓	1.03	8.7
Depression Severity → Diabetes	Logistic	14,328 / 14,418	✓	✓	✓	1.00	8.4
Uric Acid → Mortality in Diabetes	Cox	7,101 / 6,832	✓	✓	✓	1.03	14.2
Stress Hyperglycemia Ratio → Mortality in Diabetes	Cox	11,160 / 10,827	✓	✗	✓	1.12	8.3
Red Blood Cell Width → Mortality in Diabetes	Cox	3,061 / 3,157	✓	✓	✓	1.28	5.9

Table 4.2.1. Summary of the concordance evaluation from the reproduction study. Concordance was assessed across multiple metrics, including effect direction (whether both analyses indicated increased or decreased risk), p-value concordance (whether both results were statistically significant at $p < 0.05$ or both were non-significant), and overlap of 95% confidence intervals (CIs). A check mark indicates concordance, whereas a cross indicates discordance. Additional descriptive measures were used to support interpretation, including comparison of analytic sample sizes between the LATCH reproduction and the original published studies to assess cohort alignment, and calculation of relative effect size ratios (published estimate divided by the LATCH estimate). P-value concordance, confidence interval overlap, and relative effect size comparisons were based on the most extreme exposure level (for example, highest quartile or highest tertile). This table presents concordance results for logistic and Cox regression analyses, which constituted the majority of reproduced models, and is limited to these models because other analysis types reported heterogeneous outcome metrics that could not be consistently summarized within a single table. The reported time reflects the elapsed duration from query initiation to report of results. Reported runtimes are approximate and may vary with network conditions and token usage, which can differ slightly across runs even for identical queries. Additional information for all 20 studies, including references to the original publications and Figure 4.2.2, are provided in Table 4.2.2. Reproduced from ref. 22 under a CC BY 4.0 license.

Analysis	Figure	Ref	Analysis Type	Reported Values	Reproduction Assessment Bases	Concordance
Monocyte to Lymphocyte Ratio → Diabetic Retinopathy	Fig. 2a	130	Logistic Regression	OR, 95% CI, p value	Effect Direction, CI Overlap, P value significance,	Yes
Albumin → Diabetic Retinopathy		131				Yes
HDL Cholesterol → Diabetic Retinopathy		132				Partial
Lactate Dehydrogenase → Diabetic Retinopathy		133				Yes
Hemoglobin → Diabetic Retinopathy		134				Yes
UACR/SACR Ratio → Diabetic Retinopathy		135				Yes
Systematic Inflammation Index → Diabetic Kidney Disease		125				Yes
Neutrophile to Lymphocyte Ratio → Diabetic Kidney Disease		127				Partial
Weight Adjusted Waist Index → Diabetic Kidney Disease		138				Yes
Folate → Diabetic Kidney Disease		139				Yes
Food Insecurity → Depression in Diabetes		126				Yes
Neutrophile to Lymphocyte Ratio → Depression in Diabetes		127				Yes
Depression Severity → Diabetes		142				Yes
Uric Acid → Mortality in Diabetes	Fig. 2b	143	Cox Regression	HR, 95% CI, p value	Effect Direction, CI Overlap, P value significance,	Yes
Stress Hyperglycemia Ratio → Mortality in Diabetes		144				Partial
Red Blood Cell Width → Mortality in Diabetes		145				Yes
Kidney Disease in Diabetes	Fig. 2c	136	Prevalence	%	Absolute Difference within 1%	Yes
Depression in Diabetes		140				Yes
No Retinopathy vs Diabetic Retinopathy	Fig. 2d	129	Group Comparison	Mean, Standard Deviation	Direction of between-group differences, CI Overlap	Yes
Glycohemoglobin → Cognitive Test	Fig. 2e	141	Linear Regression	B, 95% CI, p value	Effect Direction, CI Overlap, P value significance,	Partial

Table 4.2.2. Summary of the 20 studies included in the LATCH reproduction analysis. Analysis refers to the analysis name reported in Fig. 2. The Ref column indicates the corresponding publication number listed in the references. The Figure column corresponds to the subfigures in Fig. 4.2.2 representing each reproduced analysis. Reported Values are the results originally published in each study, whereas Replication Assessment Bases are the metrics used to evaluate the extent of reproduction. Concordance indicates the overall reproduction outcome: Yes, if all reproduction assessment bases were fulfilled; Partial if one or more criteria showed discrepancies; and No if none of the criteria were met. Reproduced from ref. 22 under a CC BY 4.0 license.

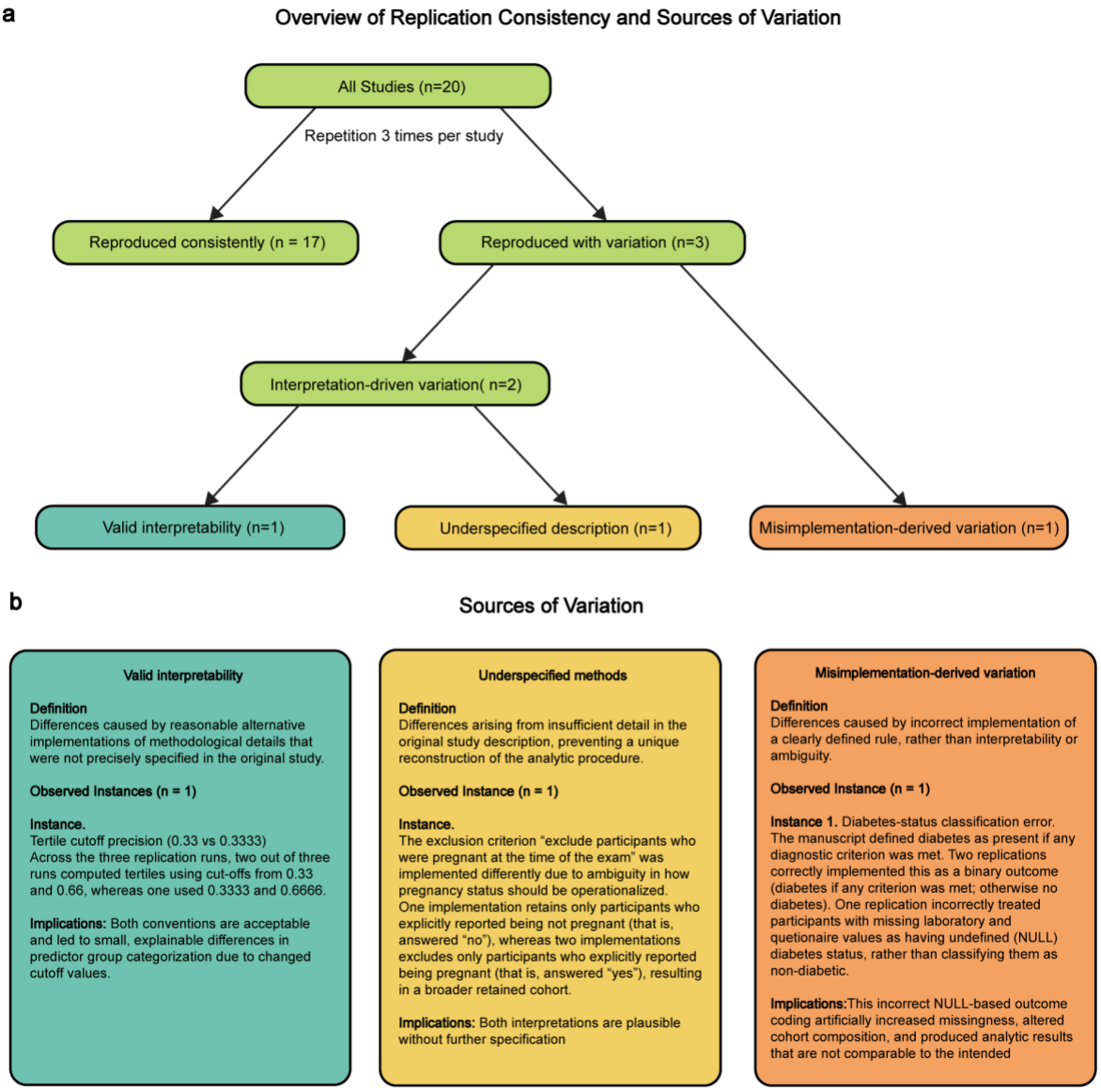


Figure 4.2.3. Repetition consistency and sources of analytic variation. a, Overview of replication outcomes across 20 studies, each repeated three times to assess consistency given the non-deterministic nature of LLM outputs. Seventeen studies reproduced consistent results, whereas three exhibited variations. Among variable cases, two were attributable to interpretation-driven differences, stemming from valid alternative implementations or underspecified descriptions, and one resulted from a misimplementation-derived error. **b,** Three illustrative examples of analytic variation, one each representing a valid interpretive difference, an underspecified method, and a misimplementation-derived variation, with definitions, observed instances, and implications for analytic reproducibility. Reproduced from ref. 22 under a CC BY 4.0 license.

4.2.4 Discussion & Conclusion

Our systematic reproduction of 20 NHANES-based diabetes studies demonstrates that LATCH can replicate published epidemiological findings to a substantial degree: 16 of 20 studies met all concordance criteria, while 4 partially met them, using only natural language descriptions of study methods (Table 4.2.2). Across logistic and Cox regression analyses, as well as descriptive and linear modeling tasks, LATCH achieved either full or partial concordance with the original publications in effect direction, statistical significance, and confidence interval overlap. Most reconstructed cohorts closely matched published sample sizes, with differences within 10%, and end-to-end analyses were completed within 3–15 minutes (Table 4.2.1). These results suggest that automated, prompt-driven analytic systems can reliably execute complex, multi-step workflows that traditionally require substantial manual effort.

The discrepancies we observed highlight persistent challenges in scientific reproducibility that extend beyond the performance of any single system. Ambiguities in cohort definitions, incomplete reporting of variable transformations, and the absence of shared analysis code complicate exact reconstruction of published studies. In several cases, reasonable alternative interpretations of underspecified methods produced measurable differences in results. Such variability underscores the importance of transparent reporting standards and machine-interpretable methodological descriptions. LATCH's structured prompting framework can serve not only as a reproduction tool but also as a lens through which to expose hidden assumptions and ambiguities in existing literature.

Our repetition experiments further reveal that, although LATCH outputs are largely stable under identical inputs, non-deterministic elements inherent to LLM-based systems can introduce occasional variability. Most variable cases reflected valid interpretive flexibility or underspecified source methods rather than systematic errors. Nevertheless, the presence of a misimplementation-derived discrepancy emphasizes the need for human oversight and verification, particularly in high-stakes biomedical contexts. Automated systems should be viewed as augmenting, rather than replacing, expert judgment.

This study has several limitations. Our curated sample prioritizes methodological clarity and feasibility of reproduction and therefore does not constitute a comprehensive survey of all NHANES-based diabetes research. We focused primarily on demographic-adjusted models to standardize comparisons, which may not capture the full complexity of fully adjusted analyses reported in original studies. Additionally, the lack of access to original analysis code prevents definitive attribution of some discrepancies. Future work could extend this framework to larger and more diverse corpora, incorporate direct code comparison when available, and evaluate performance across additional datasets.

Despite these limitations, our findings illustrate the potential of AI-assisted analytic frameworks to enhance transparency, efficiency, and reproducibility in biomedical research. By enabling rapid, standardized re-execution of published analyses, systems like LATCH can support independent verification, facilitate education and method sharing, and accelerate hypothesis testing. More broadly, integrating automated reproduction tools into the research ecosystem may encourage clearer methodological reporting and promote a culture of reproducible science.

4.3 Characterizing Robustness and Operational Limits

4.3.1 Introduction

Automated analytic systems that rely on LLMs must operate reliably under a wide range of imperfect real-world inputs. In clinical and epidemiological research, user queries may be incomplete, ambiguously phrased, semantically distant from database schemas, or logically complex. It is therefore important to characterize system limitations under edge-case conditions and to evaluate the effectiveness of internal safeguards. Understanding these operational limits is essential to ensure that AI-driven frameworks can be deployed responsibly and used with appropriate human oversight.

The reproduction study evaluated system performance under well-specified study conditions and based on details from published works, providing an assessment of functional correctness under relatively ideal inputs. In this section, we systematically probe the robustness of the LATCH framework through controlled stress testing of its LLM integrated modules: the Planner, Variable Mapper, and Data Engine. By introducing structured perturbations to validated clinical queries across multiple dimensions, including content validity, semantic variability, logical complexity, and backend model substitution, we aim to quantify failure modes, measure safeguard efficacy, and define the boundaries of reliable operation.

4.3.2 Methods

To evaluate the operational boundaries of LATCH and its robustness under edge-case conditions, we conducted a comprehensive stress-testing and safeguard assessment. System robustness was examined through controlled perturbations targeting three LLM-integrated modules: the Planner, the Variable Mapper, and the Data Engine, together with their associated safeguard mechanisms (Figure 4.3.1). Each module was evaluated independently, as failures at any stage can propagate downstream and compromise overall system performance. The evaluation was anchored on a set of validated queries derived from the three clinical domains included in the reproduction study: diabetic retinopathy, diabetic kidney disease, and depression. From the query set, one representative query per domain was selected as a baseline for systematic perturbation.

Although these baseline queries had previously demonstrated structural completeness, logical coherence, and strong alignment with the underlying database schema, they were intentionally modified to generate 400 challenging variants designed to stress different system components. Content validity testing focused on the Planner safeguard module (n = 100), evaluating the system's ability to detect missing, contradictory, or ambiguous analytical specifications from the user. Semantic variability testing targeted the Variable Mapper safeguard (n = 200), introducing diverse natural language reformulations to assess tolerance to schema misalignment and linguistic variation. Logical complexity testing focused on the Data Engine safeguard (n = 100), probing the system's capacity to preserve logical consistency and correctly translate complex analytical intent into executable SQL queries. By systematically perturbing the factual, linguistic, and logical structure of validated inputs, this evaluation framework enabled characterization of LATCH's operational limits under potentially challenging conditions.

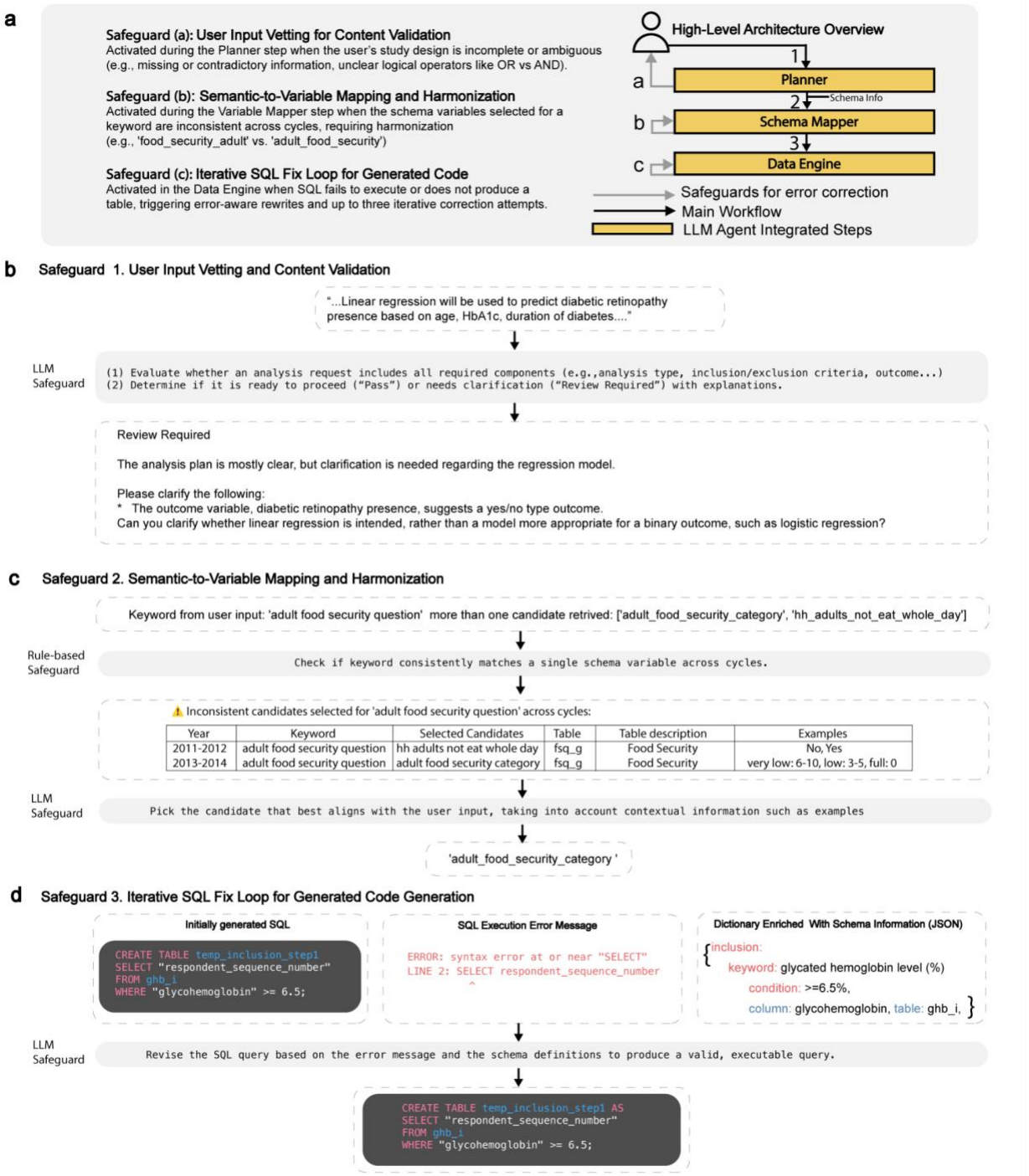


Figure 4.3.1 | Safeguards in LLM integrated modules in LATCH. a, High-level architecture showing the Planner, Schema Mapper, and Data Engine with integrated safeguards for error detection and correction. **b,** Safeguard 1 vets user input and requests clarification when specifications are not sufficient. **c,** Safeguard 2 enforces consistent semantic-to-schema variable mapping across database cycles. **d,** Safeguard 3 iteratively corrects SQL using execution feedback and schema metadata from earlier steps. Reproduced from ref. 22 under a CC BY 4.0 license.

Query Content Validity

This evaluation assessed safeguard that performs user input vetting at the Planner stage. A total of 100 queries were systematically perturbed by introducing missing critical study elements (e.g., removal of the study period or undefined outcomes), ambiguous phrasing (e.g., "exclude sick people"), or contradictory specifications (e.g., requesting linear regression for a binary outcome). Safeguard responses were classified into four outcome categories: Correct Alert (Essential Only), in which the safeguard correctly identified the issue and provided only the necessary corrective guidance; Correct Alert (With Extra Detail), in which the safeguard correctly identified the issue and included unnecessary information; No Alert (Missed Issue), in which the safeguard failed to detect the problem and proceeded with the flawed query without warning; and Wrong Alert (Irrelevant issue), in which the safeguard failed to identify the true issue and instead generated irrelevant or unnecessary feedback. Results were summarized as the percentage of queries falling into each outcome category.

Semantic Variability

This evaluation examined the robustness of the automated schema selector to natural language variation and quantified the contribution of the Variable Mapper safeguard. A systematic perturbation study was conducted in which 200 variant prompts were generated from variable phrases appearing in the validated queries from the reproduction study. Each Variant Prompt was created by applying a single-word or single-phrase substitution within the original query template using a curated synonym list for relevant variables (for example, replacing "serum albumin" with "blood albumin level").

Each Variant Prompt was processed by the Variable Mapper with the safeguard enabled, and outputs were evaluated using a binary correctness metric (1 = correct mapping, 0 = incorrect mapping), defined by whether the perturbed phrase was successfully mapped to the appropriate underlying database variable. Outputs that required safeguard intervention to achieve correct mapping were labeled Safeguard Required, whereas outputs that achieved correct mapping without intervention were labeled Safeguard Not Required.

The magnitude of semantic perturbation was quantified using a deviation score ranging from 0 to 1. Sentence embeddings were generated for the original schema phrase (A) and the perturbed variant phrase (B) using the Sentence-BERT (SBERT) framework with the all-mpnet-base-v2 model[146]. The deviation score was computed as the inverse cosine similarity:

$$\text{Deviation Score} = 1 - \text{Cosine Similarity}(A, B)$$

To isolate the effect of the internal robustness mechanism, the dataset was segmented into two groups: Safeguard Required (correctness achieved only after safeguard intervention) and Safeguard Not Required (correctness achieved without intervention). For each group, the relationship between semantic deviation (x) and the probability of successful mapping (P(Correct) = y) was modeled using a sigmoid function fitted via nonlinear least-squares optimization (scipy.optimize.curve_fit). The estimated midpoint parameter (x_0), corresponding to a 50% probability of correct mapping, was interpreted as the semantic robustness threshold. This framework enabled quantitative comparison of how the schema matching

safeguard shifted system performance under increasing semantic perturbation. Results were visualized by plotting raw binary correctness outcomes alongside the fitted sigmoid curves for each group.

Logical Complexity

This evaluation assessed the safeguard in the Data Engine step by examining the robustness of the LLM-based text-to-SQL generation module and the effectiveness of the iterative SQL correction loop under increasing logical complexity. A total of 100 queries were systematically perturbed to isolate the impact of logical structure by selectively modifying cohort selection logic and variable engineering operations. Three levels of logical complexity were defined. Easy difficulty included simple filtering and calculations. Moderate difficulty required logics that involved sequential dependencies and group-based reference computations. High difficulty involved the implementation of nested negation and hierarchical classification. System performance was measured using the final SQL execution success rate, defined as whether the generated query both executed without error and correctly implemented the intended cohort selection and variable engineering logic. Safeguard efficacy was quantified as the proportion of initially failed SQL queries that were successfully repaired within the three-attempt limit of the iterative correction loop.

API Variability

This evaluation characterized LATCH system behavior under substitution of the core large language model across all functional pipeline stages. Using the same set queries used for the 20 reproduction studies, we evaluated system performance across multiple contemporary LLM backends spanning different providers and model generations in addition to Gemini 2.5 Flash which was originally used for the reproduction study. For each provider, two representative model versions released in early and late 2025 were selected to capture generational differences in capability. The evaluated models included Claude 4.5 Sonnet (2025-09-29), Claude 3.7 Sonnet (2025-02-19), Gemini 2.0 Flash (February 2025), GPT-4.1 (2025-04-14), and GPT-5.1 (2025-11-13).

The original reproduction pipeline was developed and validated using Gemini 2.5 Flash, which served as the reference backend for the system and was therefore treated as the baseline implementation. For cross-model evaluation, the original queries corresponding to each of the 20 studies were executed without any perturbation using each alternative LLM backend. Identical prompt templates and inputs were used across all pipeline modules, including the Planner, Variable Mapper, and Data Engine, to ensure that observed differences were attributable solely to backend model substitution. Resulting system outputs were compared for correctness and reproducibility. All safeguard mechanisms were enabled during this evaluation to preserve standard operational conditions. The primary objective was to quantify end-to-end system sensitivity to backend model substitution rather than to optimize prompt performance or independently benchmark individual safeguards.

Performance was assessed using task-level correctness, defined as successful reproduction of the intended analytical logic and study outputs under identical execution conditions. To account for minor implementation-level variability that does not alter analytical meaning or cohort definition, a limited set of semantically equivalent outputs was accepted as correct. Specifically, four cases were classified as

valid despite non-identical formulations. These included: (1) use of mathematically equivalent cohort threshold expressions (for example, < versus <= when boundary values were not present in the data); (2) minor decimal precision differences in cutoff values (for example, 0.33 versus 0.3333); (3) substitution of closely related clinical variables that reflected the same clinical construct and did not alter the intended clinical interpretation; (4) logically equivalent pregnancy exclusion logic, such as explicitly excluding pregnant participants versus including only non-pregnant participants. These variations were considered analytically equivalent because they produced functionally equivalent cohort definitions and downstream results under the study data distribution. This experiment design enabled controlled comparison of system robustness across LLM providers and model generations.

4.3.3 Results

Robustness was assessed through controlled perturbations designed to probe edge-case behavior across three dimensions: Content Validity, Semantic Variability, and Logical Complexity (Fig. 4.3.2a). Each safeguard was evaluated independently, as failure at any stage can compromise overall system performance (Figure 4.3.1). A total of 400 stress-test queries were generated across the three evaluation dimensions.

In Content Validity testing, performance was strongest for detecting missing study details, with an overall accuracy of 89% (74% fully correct and 15% with extraneous information) (Fig. 4.3.2b). Performance was lower for ambiguous prompts, with correct vetting responses in 73% of cases (64% correct alerts only and 9% with extraneous information). The system failed to flag issues in 21% of cases and generated irrelevant warnings without the necessary alert in 6% of cases.

Semantic Variability testing evaluated robustness to deviations in phrasing and schema variable naming, with performance expected to decline as semantic deviation increased. For example, replacing “age in years at screening” with a low-deviation variant such as “age at screening” yielded a deviation score of 0.07, whereas a high-deviation substitution such as “time one has been around” produced a score of 0.95. Safeguard interventions improved robustness by shifting the 50% correctness threshold from 0.56 to 0.60 (Figure 4.3.2c). Remaining failures were primarily associated with high-distance substitutions, domain-specific abbreviations, and cases involving multiple closely related variables.

Logical Complexity testing categorized prompts into three tiers: Easy (simple filtering or arithmetic operations), Moderate (sequential dependencies and group-based reference computations), and Difficult (nested negation and hierarchical classification). First-pass execution success rates were 82%, 74%, and 64%, respectively, increasing to 91%, 86%, and 73% after safeguard-based corrections (Figure 4.3.2d). Safeguards improved robustness by correcting syntactic and structural errors, although nested and double-negation queries remain challenging.

To assess API-level variability, reproduction analyses from 20 published studies were repeated using LLMs from different providers and model generations. Two versions per provider, released in early and late 2025, were evaluated. The original reproductions were conducted using Gemini 2.5 Flash, which served as the reference backend during system development, and therefore near-perfect accuracy was expected for this model. Late-2025 models (GPT 5.1 and Claude 4.5 Sonnet) achieved approximately 95% correctness, followed by early-2025 models, including Claude 3.7 Sonnet (75%), GPT 4.1 (70%),

and Gemini 2.0 Flash (40%) (Figure 4.3.2e). Earlier-generation models exhibited lower performance than newer models under identical prompts and execution environments.

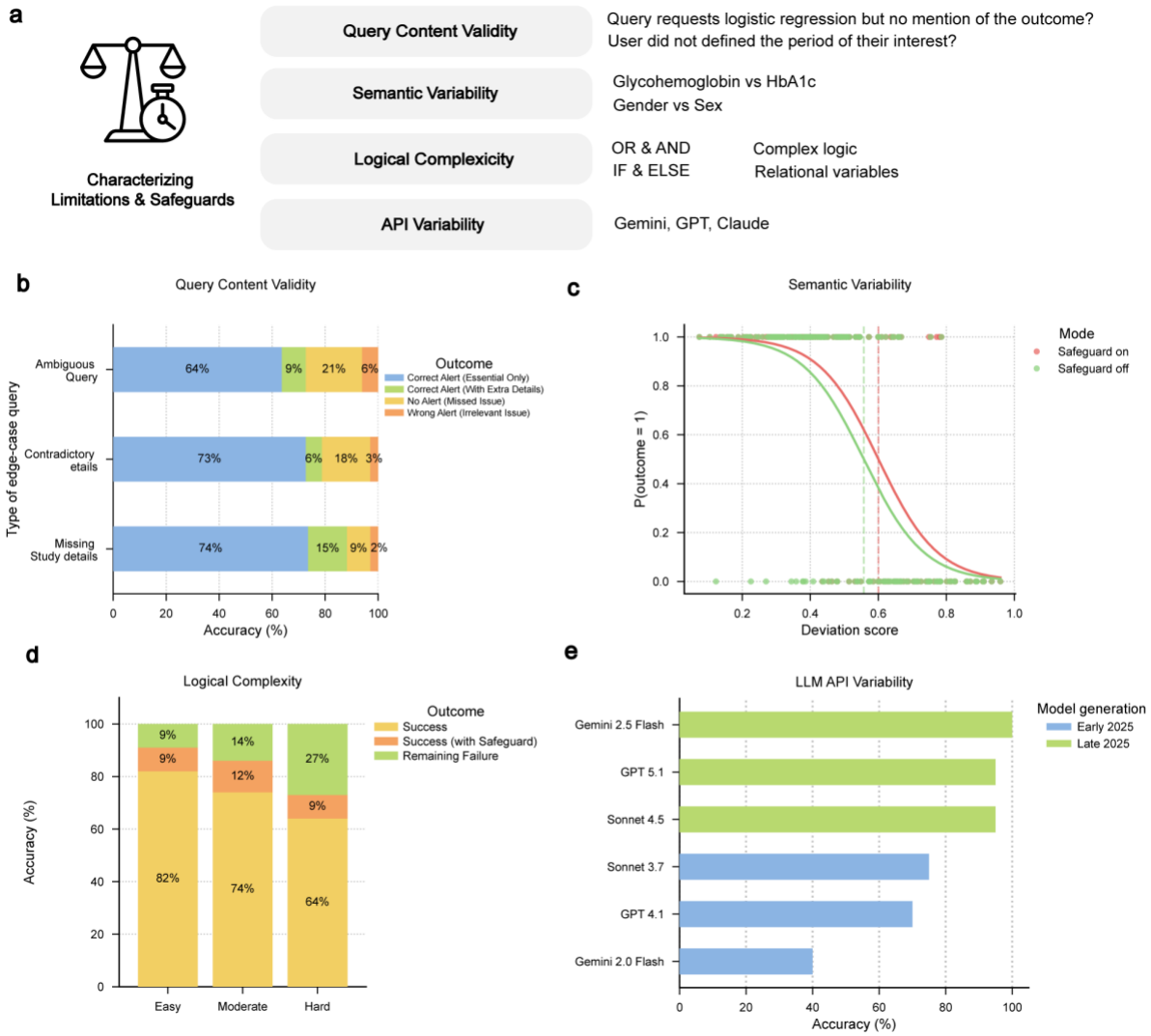


Figure 4.3.2. Characterizing limitations and safeguards of the LATCH framework. a, Overview of potential sources of failure in translating natural language clinical hypotheses into executable analyses, including query content validity (e.g., missing outcomes or contradictory study details), semantic variability (e.g., substantial variation between clinical terminology and schema variables), logical complexity (e.g., complex logic and relationally derived variables), and API-level variability across model generations. **b**, Query content validity stress test showing outcome distributions for ambiguous queries, contradictory specifications, and missing study details. **c**, Semantic variability analysis illustrating how phrase deviation from schema variables affects outcomes. Sigmoid fits demonstrate that safeguards increase tolerance to semantic deviation, shifting the estimated midpoint parameter (x_0), corresponding to a 50% probability of correct mapping, which is interpreted as the semantic robustness threshold. **d**, Logical complexity evaluation across three difficulty tiers (easy, moderate, hard). Stacked bars indicate first-pass execution success, additional corrections recovered by safeguards, and remaining failures. **e**, API-level variability across contemporary LLMs evaluated on identical queries from the reproduction study and prompts highlighting systematic performance differences by model generation. Reproduced from ref. 22 under a CC BY 4.0 license.

4.3.4 Discussion & Conclusion

Our stress-testing experiments provide a systematic characterization of LATCH's operational limits and clarify the conditions under which automated hypothesis testing remains reliable. Across content validity, semantic variability, and logical complexity evaluations, safeguard mechanisms improved system performance by detecting problematic inputs, correcting execution errors, and expanding tolerance to linguistic variation. The Planner safeguard demonstrated strong ability to identify missing or inconsistent study specifications while performing worse for catching ambiguous queries. The Variable Mapper safeguard measurably increased semantic robustness by shifting the threshold at which mapping accuracy declined. Similarly, the Data Engine safeguard recovered a fraction of initially failed SQL executions, particularly for moderately complex queries. Together, these findings indicate that layered safeguards are helpful in increasing performance in LLM-driven analytic pipelines.

At the same time, our results highlight meaningful operational limits. Performance degradation was most evident under high semantic deviation and nested logical structures. API-level variability further demonstrated that system behavior depends on the underlying language model, with newer model generations achieving substantially higher correctness than earlier versions under identical prompts. These observations suggest that robustness is not a static property of the framework alone but emerges from the interaction between system design, safeguard architecture, and the capabilities of the underlying LLM. The presence of residual failure modes reinforces the importance of human oversight, especially for complex or high-stakes analyses. Rather than serving as a fully autonomous replacement for expert analysts, LATCH is best viewed as an augmented research assistant that accelerates workflow while preserving a human expert verification step.

More broadly, this robustness evaluation establishes a template for benchmarking AI-driven scientific tools. By explicitly stress-testing edge cases and quantifying safeguard contributions, we move beyond simple accuracy metrics toward a deeper understanding of system behavior under realistic operating conditions. As AI systems become increasingly embedded in biomedical research, such systematic evaluations will be essential for building trust, guiding iterative improvement, and defining safe deployment boundaries.

Although expert human involvement remains essential, within LATCH that role is shifted more toward high-level review, methodological oversight, and adjudication than toward constructing each analytic component manually or debugging every line of code from scratch. This distinction is central to the practical value of the framework. LATCH does not remove the need for scientific judgment, but it reduces low-level implementation burden and allows investigators to focus more of their effort on study design, analytic reasoning, evaluation of methodological validity, and interpreting the numerical results. In this sense, the framework is intended to support more efficient and transparent clinical research workflows while preserving expert control over the scientific process.

Expert review remains necessary at both the level of study design and the level of stepwise analytic implementation. Investigators must ensure that the initial research question is clinically meaningful, methodologically appropriate, and sufficiently specified for analysis. This is important because when study descriptions are incomplete, ambiguous, or erroneous, as such limitations may lead to unintended analyses as shown in the content validity testing results. In addition, expert oversight is required to verify

that user intent is preserved throughout planning, variable selection, cohort definition, and code generation. Successful execution alone does not guarantee that an analysis has been implemented correctly. For example, a workflow may run without error while still misrepresenting the intended cohort or variable definitions. Safeguards can generate warnings, but these still require human adjudication and resolution. To facilitate this oversight, LATCH produces a structured analysis record that captures key implementation details, including variable mappings, code comments, and formatted workflow documentation. By presenting these elements in a transparent and organized form, the framework makes it easier for investigators to review and evaluate how the analysis was specified and executed. Human expertise is also important for interpretation of statistical outputs in clinical and epidemiologic context, including placing findings in appropriate scientific context and determining how they should be reported. Statistical results alone do not establish clinical relevance, and publication decisions often depend on considerations beyond numerical significance, such as prior literature, effect size, limitations, plausibility, and clinical relevance. Expert judgment is therefore necessary to determine whether findings are meaningful. Accordingly, LATCH should be understood as an assistive infrastructure for analytic implementation rather than as a replacement for scientific judgment.

4.4 Chapter Summary

This chapter introduced and evaluated LATCH, a privacy-aware, auditable interface for translating well-specified research questions into deterministic analyses over standardized data. The chapter first described the system architecture, which combines an LLM-assisted semantic layer with a deterministic execution layer and an analytic reporting pipeline. It then validated the framework by reproducing published diabetes studies using NHANES, demonstrating that LATCH can generate end-to-end analytical pipelines from text-based study specifications. Finally, the chapter characterized the robustness and operational limits of the framework through targeted stress testing, showing that safeguard mechanisms improve performance while also revealing important boundary conditions under which human oversight, including careful code review, remains necessary. Overall, this chapter establishes LATCH as a proof-of-concept system for reproducible and transparent clinical analytics.

Chapter 5. Application of LATCH in Advancing Knowledge

5.1 Extending Existing studies: Dataset Generalizability, Stratified Analysis, Temporal Consistency, Granular Analysis

5.1.1 Introduction

Beyond reproducing existing findings, the ultimate value of an automated analytic framework lies in its ability to accelerate new scientific discovery. Many published epidemiological studies are constrained by practical limitations, including reliance on a single dataset, arbitrarily chosen temporal windows, limited adjustment for confounding, or simplified outcome definitions. Addressing these limitations typically requires substantial manual effort to redesign analyses, harmonize datasets, and reimplement statistical workflows. As a result, potentially informative extensions of prior work are often underexplored.

In this chapter, we evaluate LATCH as a tool for advancing knowledge by enabling rapid extensions of published studies through modifications of natural language prompts. We focus on four common avenues for strengthening scientific inference: testing cross-dataset generalizability to assess external validity, examining temporal consistency to evaluate stability of associations over time, performing stratified analyses to include additional confounding and patient heterogeneity, and enhancing outcome granularity to extract higher-resolution clinical insights. By demonstrating that these extensions can be executed through minimal prompt adjustments within a unified framework, we illustrate how LATCH can transform static published findings into dynamic, extensible research projects. Portions of this chapter are adapted from the author's preprint[22].

5.1.2 Methods

The framework's utility for accelerating scientific inquiry was demonstrated by extending the findings through modifications to the original natural language queries. These extensions were designed to study the generalizability and robustness of the original associations by evaluating common limitations such as dataset specificity, temporal change, patient heterogeneity, and reliance on less granular outcomes. Specifically, the extension study comprised four distinct analyses, moving beyond the scope of the original publications: cross-dataset generalizability for testing external validity across independent cohorts; temporal consistency for assessing effect consistency across different time periods; stratified analysis with confounder adjustment; and granularity enhancement for providing higher-resolution outcome definitions.

Cross-Dataset Generalizability

To assess external validity, we evaluated previously reported risk factors for diabetic retinopathy across two independent cohorts: NHANES and AI-READI. This analysis included three risk factors from the reproduction study[131,134,135] (originally derived using NHANES data) and ten additional risk factors identified from a published American Academy of Ophthalmology (AAO)'s Diabetic Retinopathy Preferred Practice Pattern[147]. To enable a same-setting comparison, NHANES analyses were replicated within the AI-READI dataset by re-initializing the LATCH reproduction workflow with modified natural language prompts. Adjustment for demographic covariates was restricted to age, as additional demographic variables such as sex and race were not available in the publicly accessible AI-READI dataset. These prompts aligned study periods with cohort-specific data availability and defined the AI-READI "age" variable as the direct equivalent to the NHANES age variable. To facilitate a direct comparison of effect sizes, the prompts specified unweighted logistic regression models for both datasets, as the survey design variables required for weighting were not applicable for the AI-READI cohort.

Participants who had missing values in our risk factor of interest, diabetic retinopathy status or without diabetes were excluded. Most variables of interest, including BMI, hemoglobin, serum albumin, kidney disease status, glycohemoglobin, blood pressure, insulin use, LDL cholesterol, smoking history, and diabetic retinopathy status, were directly transferable across datasets through LATCH's automated variable mapping without requiring dataset-specific redefinition. However, to maintain analytical integrity, the system's automated safeguards trigger a warning when multiple potential variable matches are detected across data cycles. This occurred specifically for physical activity, where the presence of multiple overlapping NHANES questionnaire variables prompted a supervised refinement of the natural

language instruction. We explicitly specified "Physical activity (vigorous activity questionnaire)" rather than a generic "Physical activity" prompt for the NHANES analyses. This human-in-the-loop review ensured consistent variable alignment across datasets. A post hoc Bonferroni correction was applied to adjust for multiple hypothesis testing across 13 risk factors.

Temporal Consistency

To assess the stability of risk factor associations over time, we selected studies on all-cause mortality in diabetic patients with extended observation periods: one investigating the Prognostic Nutritional Index[148] and another evaluating the Weight-Adjusted Waist Index[149]. Using weighted Cox proportional hazards regression with mortality as the primary outcome, we adjusted for demographic variables across an aligned, overlapping study window from 2005 to 2018. To evaluate temporal consistency, we employed a two-tiered analytical approach consisting of a pooled analysis and a cycle-specific analysis. In the pooled analysis, a unified model was executed across the entire combined study period to establish a baseline effect size. This was complemented by a cycle-specific analysis, where the LATCH workflow was initialized to perform independent analyses for each of the seven biennial NHANES cycles within that window (2005-2006, 2007-2008, 2009-2010, 2011-2012, 2013-2014, 2015-2016, and 2017-2018). This allowed us to investigate potential fluctuations in these associations and determine whether the identified risk factors remained consistent across successive cohorts or were subject to temporal variation.

The only difference between input prompts were the study periods in the natural language input. The resulting hazard ratios and 95% confidence intervals for each cycle, comparing the highest exposure category (Q4) with the lowest (Q1), were plotted in a line graph where the Y-axis represented hazard ratio and the X-axis represented the NHANES cycle to visually assess temporal trends or consistency in effect sizes.

Stratified Analysis with Confounder Adjustment

Although serum albumin is known to be biologically associated with renal function[150,151], the initial study on the relationship between albumin and diabetic retinopathy did not account for kidney disease status.[131] To provide additional adjustment for potential confounding and to investigate patient heterogeneity in this association, we extended the original analytical framework by adding a chronic kidney disease (CKD) status as a covariate and modifying the model from standard logistic regression to stratified logistic regression. CKD was defined as an estimated glomerular filtration rate (eGFR) < 60 mL/min/1.73 m^2 or a urine albumin-to-creatinine ratio (UACR) ≥ 30 mg/g using relevant lab values. Attenuation of the serum albumin effect and changes in statistical significance were evaluated to assess confounding by kidney disease. Comparisons of effect estimates across strata were used to determine whether the association differed by underlying kidney dysfunction.

Granularity Enhancement

To provide a more detailed, stage-specific clinical characterization, we refined the analysis by moving beyond binary outcome definitions. Whereas the original reproduction study[134] categorized diabetic retinopathy (DR) as a binary outcome, we leveraged the available granular DR severity categories, including no DR, mild, moderate/severe DR, and proliferative DR. This approach allowed us to

investigate the systematic trend of different variables, including hemoglobin, duration of diabetes, glycated hemoglobin (HbA1c), red cell distribution width (RDW), across the increasing severity stages of DR. These relationships were visually presented in plots where mean lab values were plotted against the ordered DR severity levels, providing a more detailed, stage-specific profile than dichotomous comparisons.

5.1.3 Results

We used LATCH to extend prior NHANES-based studies by evaluating cross-dataset generalizability, temporal consistency, potential confounding and outcome granularity all through modifications to natural language prompts. To assess cross-dataset generalizability (Figure 5.1.1a), we modified the prompt to use the AI-READI dataset instead of NHANES. We evaluated three diabetic retinopathy risk factors reproduced from NHANES-based studies and ten additional risk factors mentioned in the American Academy of Ophthalmology (AAO) Diabetic Retinopathy Preferred Practice Pattern[147] across both NHANES and AI-READI datasets.

A high urinary albumin-to-creatinine ratio/ serum albumin-to-creatinine ratio (UACR/SACR) ratio (Q4) (AI-READI adjusted $P \leq 1.57 \times 10^{-4}$; NHANES adjusted $P \leq 4.21 \times 10^{-10}$), HbA1c ≥ 7.0% (AI-READI adjusted $P \leq 1.76 \times 10^{-2}$; NHANES adjusted $P \leq 4.79 \times 10^{-12}$), diabetes duration ≥ 10 years (AI-READI adjusted $P \leq 3.33 \times 10^{-4}$; NHANES adjusted $P \leq 4.07 \times 10^{-32}$), kidney disease presence (AI-READI adjusted $P \leq 1.56 \times 10^{-3}$; NHANES adjusted $P \leq 1.09 \times 10^{-32}$), and insulin use (AI-READI adjusted $P \leq 2.46 \times 10^{-9}$; NHANES adjusted $P \leq 1.62 \times 10^{-64}$) remained significantly associated with diabetic retinopathy in both datasets, whereas a high serum albumin (Q4) (NHANES adjusted $P \leq 1.22 \times 10^{-3}$) and hypertensive systolic blood pressure (NHANES adjusted $P \leq 6.64\times 10^{-3}$) were significantly associated only in NHANES.

To evaluate temporal consistency (Figure 5.1.1b), we adjusted the prompt to analyze seven NHANES cycles rather than a single study period. The Prognostic Nutritional Index (PNI)[148] demonstrated a consistently protective association across all survey cycles. In contrast, the Weight-Adjusted Waist Index (WWI)[149] showed substantial temporal variability, including direction reversals in the 2011−2012 cycle (OR = 0.72 (n = 872), 95% CI: 0.43, 1.20, $P \leq 2.08 \times 10^{-1}$) and the 2015−2016 cycle (OR = 0.94 (n = 1028), 95% CI: 0.32, 2.79, $P \leq 9.16 \times 10^{-1}$) , where odds ratios fell on opposite sides of the null value (OR = 1). Despite this, the pooled analysis indicated a significantly elevated risk associated with WWI.

Single-factor analyses may yield associations without sufficient clinical context[152]. To address this limitation, we used LATCH to reevaluate the reported association between low serum albumin and diabetic retinopathy[131] while accounting for an additional confounding variable, chronic kidney disease (CKD), a known cause of hypoalbuminemia and an established retinopathy risk factor [150,151] (Figure 5.1.1c). After stratifying by CKD status, a strong association was observed among individuals with CKD (OR = 0.33, 95% CI: 0.21, 0.54, $P \leq 2.80 \times 10^{-5}$), but not among those without CKD (OR = 0.75; 95% CI: 0.40, 1.40, $P \leq 3.58 \times 10^{-1}$). Lastly, we extended the hemoglobin-based analysis[134] by substituting its binary retinopathy outcome with a four-level retinopathy severity scale, allowing more granular characterization of biomarker trends across disease progression (Figure 5.1.1d).

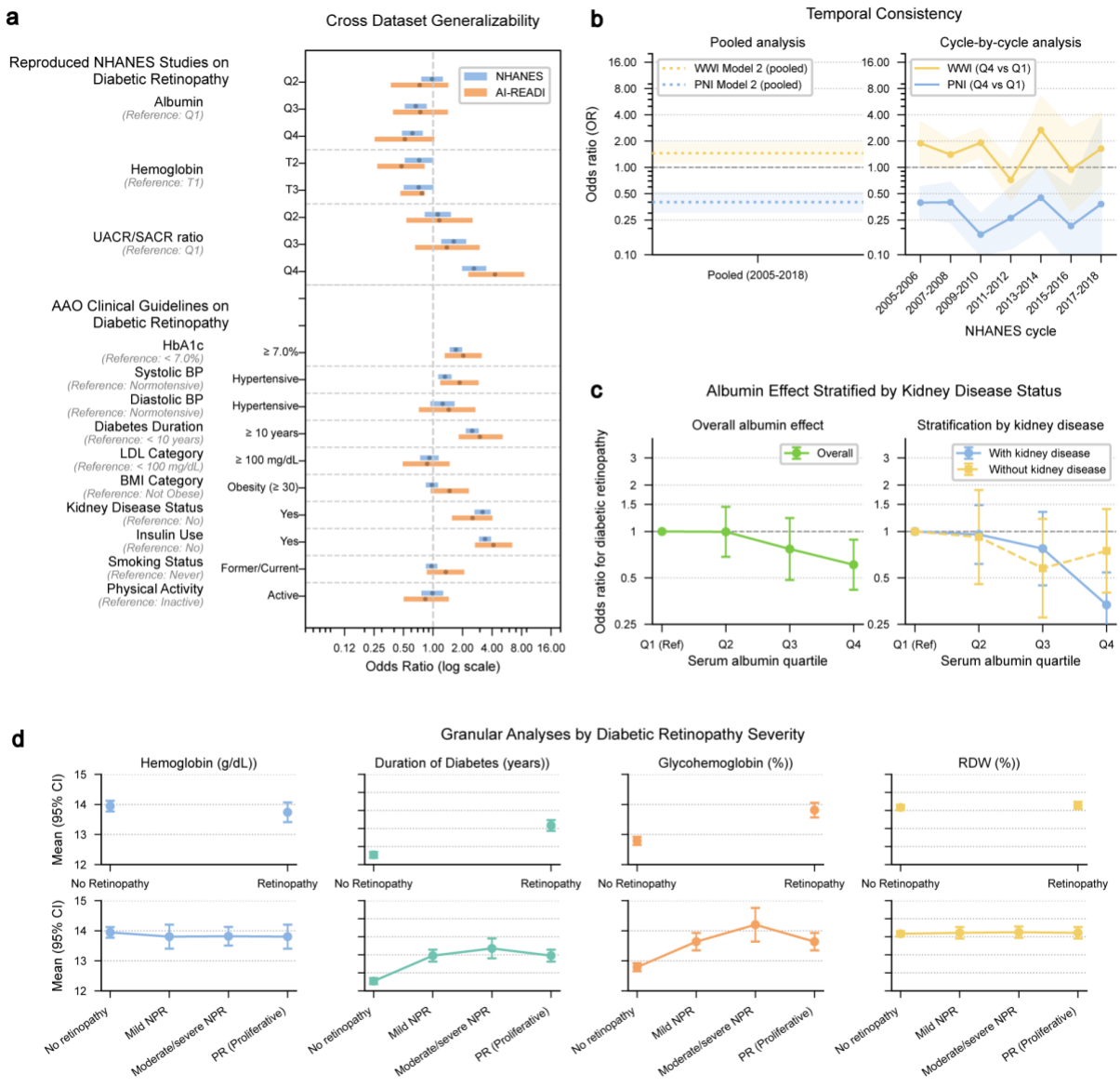


Fig 5.1.1. Extension of existing studies. Dots indicate point estimates and error bars represent 95% confidence intervals (CIs) **a,** Cross-dataset generalizability. Forest plot comparing odds ratios (ORs) for established diabetic retinopathy (DR) risk factors across the NHANES (blue) and AI-READI (orange) cohorts. Variables include risk factors reproduced from prior NHANES-based studies as well as those in the American Academy of Ophthalmology (AAO) Diabetic Retinopathy Preferred Practice Pattern. **b,** Temporal consistency. Associations between prognostic nutritional index (PNI) and mortality, and between weight-adjusted waist index (WWI) and mortality, shown for pooled analyses (left) and across individual NHANES survey cycles from 2005-2006 through 2017-2018 (right). **c,** Stratified confounder-adjusted analysis by kidney disease status. Odds ratios for DR across serum albumin quartiles shown without stratification (left) and stratified by kidney disease status (right). **d,** Granular analyses of laboratory values by diabetic retinopathy (DR) severity. The top panels replicate the original binary groups from prior studies, comparing the presence versus absence of DR. The bottom panels extend this analysis using LATCH by categorizing DR into four ordered severity levels, enabling finer-grained assessment of biomarker trends across disease progression. Dots indicate point estimates, with error bars representing 95% CIs. Reproduced from ref. 22 under a CC BY 4.0 license.

5.1.4 Discussion & Conclusion

Our extension analyses demonstrate that LATCH can function as a practical platform for moving beyond replication toward active knowledge generation. By modifying natural language prompts, we were able to systematically test external validity, temporal consistency, confounding structure, and outcome granularity across multiple clinical questions. These experiments highlight how automated analytic frameworks can lower the barrier to performing sophisticated secondary analyses that would otherwise require substantial technical effort.

Taken together, these extensions illustrate a shift from static reproduction to dynamic exploration. LATCH enables researchers to treat published analyses as starting points for iterative investigation, where alternative datasets, time windows, covariate structures, and outcome definitions can be evaluated with minimal friction. This capability has important implications for accelerating scientific discovery, improving robustness of evidence, and promoting a culture of continuous validation and refinement.

This study also highlights important considerations. Cross-dataset comparisons remain constrained by variable availability and harmonization challenges, and automated analyses do not eliminate the need for domain expertise in interpreting results. Statistical significance differences across datasets or time periods should be understood in the context of sample size, measurement error, and underlying population differences. As with any automated system, human oversight is essential to ensure that extensions remain clinically meaningful and methodologically sound.

In conclusion, LATCH demonstrates strong potential as a platform for advancing biomedical knowledge by enabling extensible exploration of published findings. By lowering technical barriers to dataset generalization, temporal validation, stratified modeling, and granular outcome analysis, automated analytic frameworks can transform how researchers interact with existing literature. Rather than viewing studies as fixed endpoints, LATCH supports an iterative model of research in which hypotheses are continuously tested, refined, and expanded. Such tools may play an important role in fostering a more dynamic and cumulative scientific ecosystem.

5.2 Hypothesis-Generating Analyses: National Level Trend in NHANES to Granular Analysis in AI-READI

5.2.1 Introduction

In addition to extending prior studies, we used LATCH to conduct hypothesis-generating analyses examining the associations between diabetes severity and visual function and retinal structure. Connecting broad epidemiological trends in population-level prevalence from NHANES with granular functional and structural measurements from the AI-READI dataset provides a more comprehensive view of how diabetes and vision are related.

In this section, we apply a sequential, hypothesis-driven workflow in which each analysis informs the next. We begin with population level trend identification in NHANES to examine whether diabetes prevalence differs across vision impairment categories. We then use the AI-READI cohort, which

contains more detailed visual acuity measures, retinal imaging, and granular measures of diabetes severity, to test more specific hypotheses about visual function deficits and structural retinal changes associated with diabetes severity. Finally, we evaluate conditional associations between retinal structure and visual performance to assess whether structural alterations statistically account for observed functional differences. The aim of this chapter is to demonstrate how LATCH can support iterative exploratory, hypothesis-generating analyses across datasets and analytical scales while maintaining transparent and reproducible analytical specifications. The analyses in this section should be interpreted as hypothesis-generating rather than confirmatory. Given the risks of overinterpretation and multiple testing in exploratory settings, the observed patterns and associations should be considered preliminary findings requiring independent validation. Portions of this chapter are adapted from the author's preprint[22].

5.2.2 Methods

We used LATCH to conduct hypothesis-generating analyses through a sequential workflow. Analyses were performed using natural language specifications, enabling automated end-to-end execution while maintaining analytical transparency. We first examined diabetes prevalence across vision impairment groups in NHANES and then extended these analyses to the AI-READI cohort, which provides detailed ophthalmic phenotyping and retinal structural measures. All steps, including data selection, variable definitions, exclusion criteria, and statistical modeling, were defined via natural language prompts.

Trend Identification (NHANES)

For population-level trend identification, we used NHANES data from five survey cycles (1999-2008) which reported vision data. We included participants aged 20 years or older with complete data for both presenting and post-refraction visual acuity in both eyes, and we excluded individuals missing diabetes related information, defined as absence of both glycohemoglobin (HbA1c) measurements and self-reported physician diagnosis of diabetes. Diabetes was a binary variable, defined as HbA1c $\geq$ 6.5% or clinician-reported diabetes. Visual status was derived using the better-seeing eye and categorized into a three-level group variable: no visual impairment (20/40 or better), correctable visual impairment (presenting acuity worse than 20/40 but post-refraction acuity 20/40 or better), and nonrefractive visual impairment (worse than 20/40 even after refraction). Using NHANES survey weights, exam weights, and accounting for the complex sampling design, we estimated weighted prevalence of diabetes across visual impairment groups and survey cycles to assess population-level trends in vision-associated diabetes burden.

Visual Deficit Analysis (AI-READI)

To investigate functional visual deficits associated with diabetes more precisely, we performed linear regression analyses using the AI-READI dataset collected during 2023-2025. Participants were excluded if data were missing for the diabetes severity group or for required visual outcome and covariate variables. Three visual outcomes were analyzed. Photopic vision was defined using photopic logMAR acuity as the better (lower) value of the two eyes, multiplied by -1 so that higher scores indicate better vision. Mesopic vision was defined analogously to day vision, using mesopic logMAR values with the

lower value multiplied by -1. Contrast sensitivity was defined as the higher value between the two eyes using log contrast sensitivity measurements. Both photopic and mesopic contrast sensitivity data was used. In all visual function models, the primary predictor was study group, representing diabetes severity, and age and ocular comorbidities (reported diagnosis of glaucoma, cataract, diabetic retinopathy, age-related macular degeneration, and retinal vascular occlusion) were included as covariates. A post hoc Bonferroni correction was applied to adjust for multiple hypothesis testing across 4 vision metrics.

Structural Retinal Layer Thickness Analysis (AI-READI)

To characterize structural and abnormalities in diabetes, we analyzed retinal layer thicknesses in micrometers extracted from OCT scans, including retinal nerve fiber layer (RNFL), ganglion cell layer (GCL), inner plexiform layer (IPL), inner nuclear layer (INL), outer plexiform layer to external limiting membrane (OPL-ELM), photoreceptor inner segment layer (external limiting membrane to ellipsoid zone; ELM-EZ), photoreceptor outer segment layer (ellipsoid zone to retinal pigment epithelium; EZ-RPE), and retinal pigment epithelium layer (RPE+Bruch's). Participants were excluded if data were missing for the diabetes severity group or for required retinal layer information and covariate variables. For each layer, we fit a separate linear regression model with study group as the main predictor while adjusting for age and ocular comorbidities (reported diagnosis of glaucoma, cataract, diabetic retinopathy, age-related macular degeneration, and retinal vascular occlusion). This allowed us to identify layers that are selectively affected in diabetes, providing structural correlates to the observed functional deficits. A post hoc Bonferroni correction was applied to adjust for multiple hypothesis testing across 8 retinal layers.

Conditional Association analysis (AI-READI)

Among significant associations identified in the structural retinal layer thickness analysis (RNFL, ELM-EZ, EZ-RPE, and RPE+Bruch's), we conducted a mediation-style conditional association analysis using the LATCH mediation module to evaluate whether retinal layer thickness statistically accounted for the association between diabetes severity and vision outcomes. Given the cross-sectional nature of the data and lack of temporal ordering, results were interpreted as conditional associations rather than causal mediation effects. Participants with complete data on diabetes severity (X), retinal layer thickness (M), vision outcomes (Y), and age and ocular covariates were included. Using ordinary least squares (OLS) regression in R, we fit sequential models and estimated indirect and direct association components using the mediate function from the mediation R package, as implemented within the LATCH mediation module. Sequential models were specified as follows:

$$\text{Model 1: } M \sim X + \text{covariates}$$

$$\text{Model 2: } Y \sim X + M + \text{covariates}$$

$$\text{Model: 3 } Y \sim X + \text{covariates}$$

Regression coefficients were summarized as path a (association between X and M from Model 1, path b (association between M and Y from Model 2, and c′ (conditional association between X and Y from Model 2). The total association (c) was obtained from Model 3. The indirect association was quantified as the product a × b, representing the attenuation of the total diabetes-vision association when accounting for

retinal layer thickness. This attenuation was reported as an absolute effect size (β), expressed in the original outcome scale (for example, log contrast sensitivity units). In addition, the proportion of the total association between diabetes severity and vision outcomes explained by retinal thickness was calculated as (a × b) / c and reported as a percentage. Confidence intervals for the indirect association (a × b) were estimated using bootstrap resampling, implemented through the mediate function. A post hoc Bonferroni correction was applied to adjust for multiple hypothesis testing across 16 analyses derived from four retinal layers and four vision metrics. For visualization, we plotted only retinal layer-vision pairs in which both the total diabetes severity-vision association and the attenuation after conditioning on retinal thickness were statistically significant after correction, and where the direction of the diabetes effect was consistent with reduced visual function.

5.2.3 Results

Beyond replicating prior studies, we used LATCH to conduct hypothesis-generating analyses through a sequential, hypothesis-driven workflow. Focusing on diabetes and vision outcomes, we used LATCH to execute analyses expressed in natural language, with each analysis motivated by the previous one. We first identified population-level patterns in diabetes prevalence across vision impairment groups in NHANES, then extended these findings to the AI-READI cohort, which provides richer vision phenotyping and retinal structural measures (Fig. 5.2.1a). This iterative process was enabled by LATCH's ability to reduce technical overhead while preserving transparency and analytical control.

In NHANES, survey-weighted group comparisons using the Rao–Scott chi-square test showed a significant association between visual impairment categories and diabetes prevalence ($P \leq 2.91 \times 10^{-2}$). The weighted prevalence of diabetes was highest among individuals with irreversible vision impairment (20.31%, 95% CI: 15.85, 24.79), compared with those with correctable impairment (13.32%, 95% CI: 10.99,15.66) or no impairment (14.64%, 95% CI: 12.59, 16.68) (Fig. 5.2.1b).

To further characterize this pattern, we utilized the AI-READI dataset, which provides more visual function metrics and granular diabetes severity annotations. Linear regression revealed lower visual function across vision measures with increasing diabetes severity (Fig. 5.2.1c). Both insulin-treated and oral medication-treated groups showed significantly reduced photopic contrast sensitivity (log; n = 2200) (insulin: β = -0.064, 95% CI: -0.085, -0.044, adjusted $P \leq 2.47 \times 10^{-9}$; oral medication: β = -0.025, 95% CI: -0.039, -0.011, adjusted $P \leq 2.32 \times 10^{-10}$) and mesopic contrast sensitivity (log; n = 2194) (insulin: β = -0.085, 95% CI: -0.110, -0.060, adjusted $P \leq 1.60 \times 10^{-10}$; oral medication: β = -0.023, 95% CI: -0.040, -0.005), adjusted $P \leq 4.52 \times 10^{-2}$). In addition to contrast sensitivity measures, both reduced photopic vision and mesopic vision were observed in the insulin group. Among these effects, the largest magnitude reduction was observed for mesopic contrast sensitivity in the insulin-treated group.

Associations between retinal structural thickness and diabetes severity were quantified using optical coherence tomography (OCT) across eight retinal layers (Fig. 5.2.1d). Significant thinning was observed in the retinal nerve fiber layer (RNFL), photoreceptor inner segment layer (external limiting membrane to ellipsoid zone; ELM-EZ), and photoreceptor outer segment layer (ellipsoid zone to retinal pigment epithelium; EZ-RPE) in both the oral medication-treated and insulin-treated groups (Figure 5.2.1e). The largest effect was the thinning of the photoreceptor outer segment (EZ-RPE, μm; n = 1,889), which was

evident in the oral medication group and exacerbated in the insulin-dependent group (insulin: $\beta = -1.58$, 95%CI: -2.01 to -1.15, adjusted $P \leq 9.59 \times 10^{-12}$; oral medication: $\beta = -0.52$, 95% CI: -0.83 to -0.22, adjusted $P \leq 5.35 \times 10^{-3}$).

To investigate the association between retinal structure and functional deficits, we performed a conditional association analysis across diabetes severity and vision metrics. This analysis quantified the association between diabetes severity and vision loss after accounting for retinal layer thickness. While multiple associations were significantly reduced after conditioning on retinal thickness (Figure 5.2.1f), the most pronounced finding involved EZ-RPE layer thickness and mesopic contrast sensitivity in the insulin-dependent group relative to healthy participants.

Compared to healthy participants, the insulin-dependent group showed the unadjusted association with mesopic contrast sensitivity of ($\beta = -0.086$, 95% CI: -0.113, -0.059, adjusted $P \leq 1 \times 10^{-3}$). Conditioning on EZ-RPE thickness yielded a significant attenuation ($\beta = -0.015$, 95% CI: -0.020, -0.010, adjusted $P \leq 1 \times 10^{-3}$), indicating that structural thinning of the EZ-RPE layer statistically accounted for 17.31% of the difference in mesopic contrast sensitivity between insulin-dependent and healthy participants. Although these cross-sectional findings reflect statistical associations rather than causal mechanisms, they suggest that thinning of the photoreceptor complex is a measurable structural correlate of functional impairment in advanced diabetes.

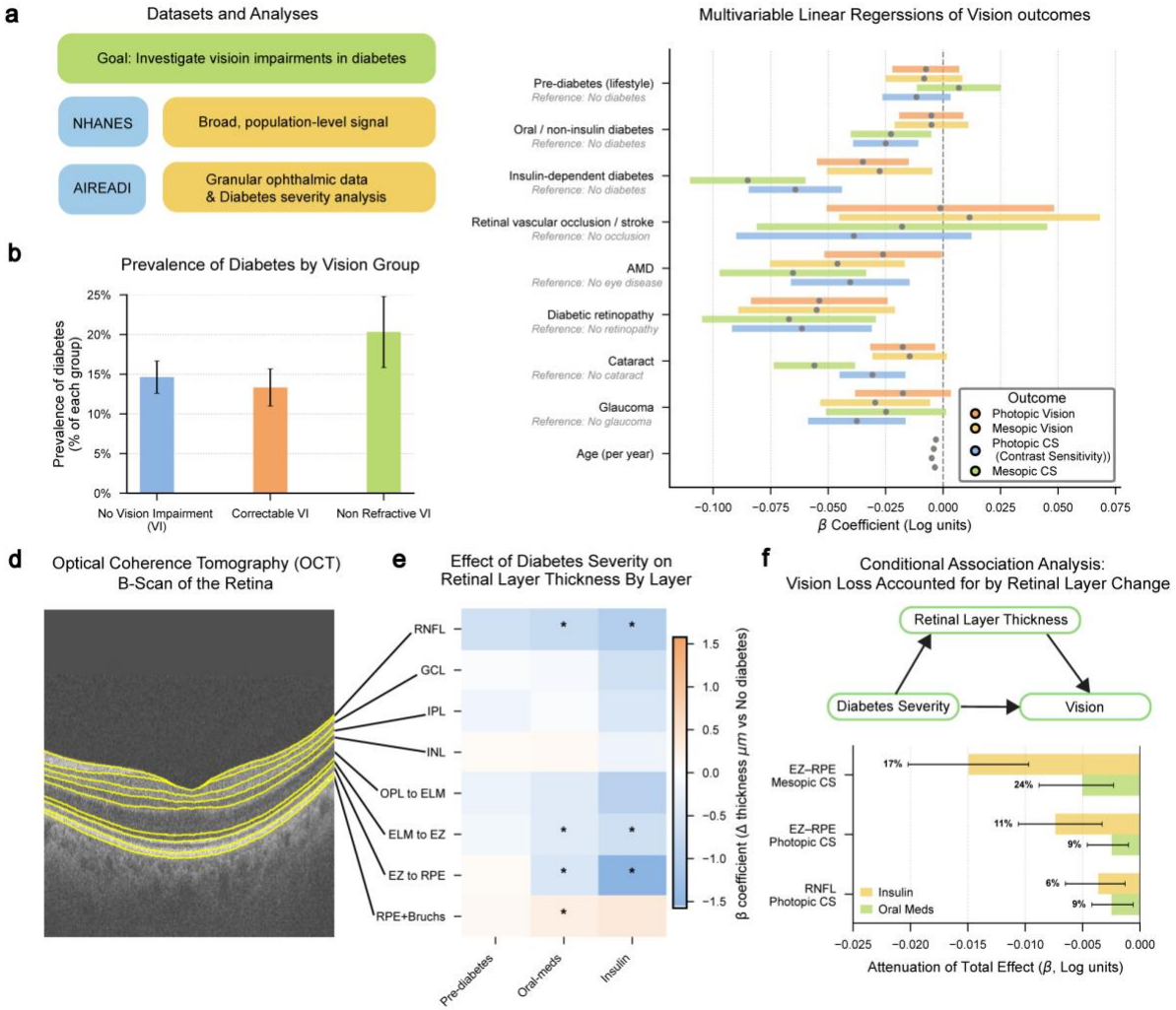


Figure 5.2.1. Hypothesis-generating analyses of diabetes, visual function, and retinal structure. a, A schematic illustrating population-level analyses, followed by granular ophthalmic phenotyping to investigate visual impairments in diabetes. **b,** Population-Level Association. Analysis reveals a significant difference in diabetes prevalence across visual impairment groups: no vision impairment (VI), correctable VI, and non-refractive (uncorrectable) VI ($P \leq 2.91 \times 10^{-2}$) from a survey-weighted group comparison using the Rao–Scott chi-square test. The highest prevalence of diabetes was observed among individuals with non-refractive VI. Error bars indicate 95% confidence intervals (CIs). **c,** Visual Function Analysis. linear regression coefficients from four separate models assessing the association between diabetes severity and visual function outcomes, including photopic and mesopic vision (−logMAR) and contrast sensitivity (log). Larger coefficients indicate better visual performance. Dots indicate point estimates, with error bars representing 95% CIs. **d,** A Schematic of the retinal layers segmented from OCT scans and their names listed from the inner to the outer retina: the Retinal Nerve Fiber Layer (RNFL), Ganglion Cell Layer (GCL), Inner Plexiform Layer (IPL), Inner Nuclear Layer (INL), the complex from the Outer Plexiform Layer to the External Limiting Membrane (OPL to ELM), ELM to Ellipsoid Zone (ELM to EZ), EZ to Retinal Pigment Epithelium (EZ to RPE), and the Retinal Pigment Epithelium + Bruch's Membrane complex (RPE + Bruchs). **e,** Retinal Layer thickness analysis. The heatmap visualizes the coefficients from eight linear regression models that tested the association between diabetes severity and the thickness of a specific retinal layer, highlighting layer-specific changes in different severity groups. Because CIs cannot be directly visualized within the heatmap, asterisks indicate significant coefficients (adjusted $P < 0.05$) after Bonferroni correction. **f,** Conditional Association Analysis. Among significant associations identified in the retinal layer thickness analysis, we evaluated the association between diabetes severity and vision outcomes after conditioning on layer thickness, across four vision measures**.** x-axis shows attenuation of the diabetes-vision association after accounting for retinal thickness, with percentages showing the proportion of the total association accounted for. Error bars represent 95% CIs. The y-axis lists retinal layer-vision outcome pairs with Bonferroni-significant total associations and significant attenuation after conditioning. Reproduced from ref. 22 under a CC BY 4.0 license

5.2.4 Discussion & Conclusion

This sequential analysis combines population-level patterns with detailed ophthalmic phenotyping to test previously unexplored hypotheses about the relationship between diabetes, visual function, and retinal structure within the analyzed datasets. Using NHANES, we observed that diabetes prevalence was highest among individuals with nonrefractive visual impairment, suggesting a disproportionate burden of diabetes in groups with persistent vision deficits. While this association does not establish directionality, it motivates closer examination of functional and structural vision changes in more characterized cohorts.

Analyses in the AI-READI dataset showed reductions in visual performance with increasing diabetes severity, particularly for contrast sensitivity under mesopic conditions. Structural analyses further identified selective thinning in specific retinal layers, most prominently within photoreceptor-associated regions. The magnitude and consistency of these layer-specific changes suggest that diabetes may differentially affect outer retinal structures.

Conditional association analyses indicated that thinning of the photoreceptor outer segment layer (EZ-RPE) statistically accounted for a measurable portion of the association between advanced diabetes and reduced mesopic contrast sensitivity. Although these cross-sectional results cannot establish causal mediation, they suggest that structural alterations in the photoreceptor complex are associated with functional deficits among individuals with advanced diabetes. This relationship offers a plausible structural correlate for the observed vision changes and suggests a direction for future longitudinal and mechanistic studies.

Several limitations should be considered. The analyses are cross-sectional and therefore cannot determine temporal ordering or causality. Differences between NHANES and AI-READI in sampling design and measurement protocols may influence comparability. In addition, given the breadth of hypotheses that can be examined within this framework and the potential for multiple comparisons, these findings should be interpreted as hypothesis-generating unless independently validated.

Overall, these results illustrate how a sequential, hypothesis-driven workflow can be used to connect epidemiological observations with detailed phenotypic measurements. By enabling iteration across datasets and analytical levels, LATCH supports the identification of new associations through exploratory, hypothesis-generating analyses. Future work incorporating longitudinal data and experimental validation will be necessary to clarify the mechanisms underlying the observed structural-functional relationships in diabetes-related vision impairment.

5.3 Chapter Summary

This chapter demonstrated the value of LATCH as a tool not only for reproducing prior work, but also for advancing biomedical knowledge. First, it showed that published studies can be extended efficiently through simple modifications of natural language prompts, enabling analyses of cross-dataset generalizability, temporal consistency, stratified subgroups, and more granular outcomes. Second, it illustrated how LATCH can be used to test previously unexplored hypotheses and conduct hypothesis-generating analyses from multimodal data, including nationwide trends in vision-related outcomes and

associations between retinal biomarkers and disease severity. These results highlight the broader scientific utility of the framework by showing that automated, transparent workflows can support both confirmation of prior evidence and discovery-focused hypothesis testing.

Chapter 6. Discussion & Future Work

Studies in this thesis demonstrate that an LLM-assisted framework can accelerate the testing of clinical hypotheses from large health datasets. We show that LATCH can be used to reproduce and extend established analyses and conduct hypothesis-generating analyses through natural language, transforming workflows that typically require time-intensive manual coding into verifiable analyses executed in minutes. Together, these findings suggest a practical step toward more efficient, accessible data-driven medical research by reducing technical and resource barriers to hypothesis testing.

Beyond efficiency, this framework enforces a higher standard of transparency and reproducibility. Reproduction of 20 published analyses exposed methodological ambiguity, where implementation details were often under-specified. Encoding this information within a LATCH prompt and its Analytic Report yields transparent, computationally reproducible analyses.

However, the successful deployment of LATCH in this study was contingent on the use of standardized cross sectional tabular datasets like NHANES and AI-READI. We recognize that this represents an idealized scenario compared to the complexity of "real-world" EHRs, which are notoriously heterogeneous[153,154]. Applying this framework to such environments is a significant challenge and a critical direction for future work as the present findings may overestimate performance in operational clinical environments. Yet this limitation also highlights an opportunity. Our findings suggest a substantial return on investment from data standardization efforts such as the Observational Medical Outcomes Partnership Common Data Model[155]. While standardizing data requires a significant one-time effort, it enables the scalability of evidence with tools like LATCH.

An additional limitation is the inherent ambiguity of natural language as it may produce various valid interpretations, highlighting the critical role of human oversight in ensuring alignment with researcher intent. Another limitation stems from the stochastic nature of LLM inference. Although LATCH produces consistent outputs under well-specified prompts, model responses remain probabilistic, and commercially hosted LLMs may undergo undisclosed architectural or inference-level updates that alter behavior over time. These may limit strict reproducibility and highlight the importance of version tracking and record-keeping.

A related consideration is the potential for rapid, automated analyses, which increase the risk of false-positive findings and may lower barriers to uncontrolled exploratory testing. Although LATCH provides an auditable record of analytic steps, it does not eliminate risks associated with multiple testing or selective reporting, underscoring the need for prespecification of outcomes and adherence to statistical principles. These risks can be mitigated by requiring external validation of findings in independent datasets, preregistration, and transparent tracking of hypotheses through registries that record all attempted analyses and promote appropriate correction for multiple testing. As LLM-based analytic tools

continue to evolve, such challenges are likely to become unavoidable and proactively addressing them through deliberate methodological standards will be essential.

Extending LATCH beyond structured tabular data toward richer multimodal clinical information is a natural next step, such as imaging-derived features transformed into tabular representations, as demonstrated with retinal thickness analysis. Future work may explore the development of a fully multimodal LATCH framework capable of interfacing directly with popular pretrained image foundation models through lightweight processes. Rather than retraining modality-specific systems from scratch, LATCH could leverage advances in large-scale visual models for inference while preserving analytic transparency and modularity and use the features from inference as analyses.

Another important extension is the incorporation of data that more closely reflects real-world clinical environments in two complementary directions. First, future work should emphasize richer longitudinal time-series data, as the current study includes only limited temporal mortality information. Expanding to high-resolution time-series data would better approximate the complexity of electronic health records and enable evaluation of LATCH in settings that require temporal reasoning and longitudinal analysis. However, working with such longitudinal clinical data is typically far more time-consuming and methodologically complex, as it involves messy, irregular sampling, missingness, and heterogeneous data sources. Addressing these challenges will require additional methodological development in data preprocessing, harmonization, and robust temporal modeling. Furthermore, integrating unstructured clinical data through natural language processing and other machine learning models represents a critical step toward more realistic multimodal analytics. Features extracted from free-text notes, or other non-tabular modalities can be incorporated into structured analytic pipelines as derived tabular variables. As this framework scales, a practical and maintainable strategy is to preserve raw data as a ground-truth source of record while developing modular feature-extraction pipelines that generate structured representations for analysis. Such an architecture supports scalability, maintainability, and reproducibility, as feature extraction is an unavoidable component of most statistical analyses and can evolve independently without altering the underlying data source.

As an LLM-assisted framework that manages the end-to-end pipeline from a natural language question to a verifiable statistical result, LATCH offers a practical vision for the future of biomedical research. It shifts scientific effort from repetitive coding toward hypothesis formulation, analysis evaluation, and result interpretation, emphasizing researcher focus on insight generation, review, and validation rather than manual implementation. Reproducibility becomes an intrinsic property of the analytic workflow rather than a retrospective concern. Together, these advances point toward a more efficient, transparent, and scalable ecosystem for translating clinical data into actionable knowledge, with the potential to accelerate scientific discovery and its impact on human health.

Chapter 7. Bibliography

1. Kahn, M. G. & Weng, C. Clinical research informatics: a conceptual perspective. *J. Am. Med. Inform. Assoc.* **19**, e36–42 (2012).
2. Richesson, R. L. *et al.* Electronic health records based phenotyping in next-generation clinical trials: a perspective from the NIH Health Care Systems Collaboratory. *J. Am. Med. Inform. Assoc.* **20**, e226–31 (2013).
3. Hripcsak, G. & Albers, D. J. Next-generation phenotyping of electronic health records. *J. Am. Med. Inform. Assoc.* **20**, 117–121 (2013).
4. Baker, M. 1,500 scientists lift the lid on reproducibility. *Nature* **533**, 452–454 (2016).
5. Niven, D. J. *et al.* Reproducibility of clinical research in critical care: a scoping review. *BMC Med.* **16**, 26 (2018).
6. Peng, R. D. Reproducible research in computational science. *Science* **334**, 1226–1227 (2011).
7. Cobey, K. D. *et al.* Biomedical researchers' perspectives on the reproducibility of research. *PLoS Biol.* **22**, e3002870 (2024).
8. Kagawa, R. *et al.* Development of type 2 diabetes mellitus phenotyping framework using expert knowledge and machine learning approach. *J. Diabetes Sci. Technol.* **11**, 791–799 (2017).
9. Banda, J. M., Seneviratne, M., Hernandez-Boussard, T. & Shah, N. H. Advances in electronic phenotyping: From rule-based definitions to machine learning models. *Annu. Rev. Biomed. Data Sci.* **1**, 53–68 (2018).
10. Van Veen, D. *et al.* Adapted large language models can outperform medical experts in clinical text summarization. *Nat. Med.* **30**, 1134–1142 (2024).
11. Croxford, E. *et al.* Evaluating clinical AI summaries with large language models as judges. *NPJ Digit. Med.* **8**, 640 (2025).
12. Johri, S. *et al.* An evaluation framework for clinical use of large language models in patient interaction tasks. *Nat. Med.* **31**, 77–86 (2025).
13. Tao, X. *et al.* An LLM chatbot to facilitate primary-to-specialist care transitions: a randomized

controlled trial. *Nat. Med.* 1–9 (2026).

14. Hager, P. *et al.* Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nat. Med.* **30**, 2613–2622 (2024).
15. Sandmann, S. *et al.* Benchmark evaluation of DeepSeek large language models in clinical decision-making. *Nat. Med.* **31**, 2546–2549 (2025).
16. Comeau, D. S., Bitterman, D. S. & Celi, L. A. Preventing unrestricted and unmonitored AI experimentation in healthcare through transparency and accountability. *NPJ Digit. Med.* **8**, 42 (2025).
17. Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* **1**, 206–215 (2019).
18. Wiens, J. *et al.* Do no harm: a roadmap for responsible machine learning for health care. *Nat. Med.* **25**, 1337–1340 (2019).
19. Poon, A. I. F. & Sung, J. J. Y. Opening the black box of AI-Medicine. *J. Gastroenterol. Hepatol.* **36**, 581–584 (2021).
20. Dennstädt, F., Hastings, J., Putora, P. M., Schmerder, M. & Cihoric, N. Implementing large language models in healthcare while balancing control, collaboration, costs and security. *NPJ Digit. Med.* **8**, 143 (2025).
21. Jonnagaddala, J. & Wong, Z. S.-Y. Privacy preserving strategies for electronic health records in the era of large language models. *NPJ Digit. Med.* **8**, 34 (2025).
22. Gim, N. *et al.* An LLM-assisted framework for accelerated and verifiable clinical hypothesis testing from electronic health records. *medRxiv* (2026) doi:10.64898/2026.02.10.26346008.
23. Gim, N., Ferguson, A. N., Blazes, M., Lee, C. S. & Lee, A. Y. The march to harmonized imaging standards for retinal imaging. *Prog. Retin. Eye Res.* **107**, 101363 (2025).
24. Current Edition. *DICOM* https://www.dicomstandard.org/current.
25. Data Standardization – OHDSI. https://www.ohdsi.org/data-standardization/.
26. Bommasani, R. *et al.* On the opportunities and risks of foundation models. *arXiv [cs.LG]* (2021).

27. Vaswani, A. *et al.* Attention is All you Need. *Advances in Neural Information Processing Systems* **30**, (2017).
28. Brown, T. B. *et al.* Language Models are Few-Shot Learners. *arXiv [cs.CL]* (2020).
29. Boiko, D. A., MacKnight, R., Kline, B. & Gomes, G. Autonomous chemical research with large language models. *Nature* **624**, 570–578 (2023).
30. Majumder, B. P. *et al.* Data-driven discovery with large generative models. *arXiv [cs.CL]* (2024).
31. Ifargan, T., Hafner, L., Kern, M., Alcalay, O. & Kishony, R. Autonomous LLM-driven research from data to human-verifiable research papers. *arXiv [q-bio.OT]* (2024) doi:10.48550/arXiv.2404.17605.
32. Scherbakov, D., Hubig, N., Jansari, V., Bakumenko, A. & Lenert, L. A. The emergence of large language models as tools in literature reviews: a large language model-assisted systematic review. *J. Am. Med. Inform. Assoc.* **32**, 1071–1086 (2025).
33. Cai, X. *et al.* Utilizing Large language models to select literature for meta-analysis shows workload reduction while maintaining a similar recall level as manual curation. *BMC Med. Res. Methodol.* **25**, 116 (2025).
34. LitLLM: A Toolkit for Scientific Literature Review. https://arxiv.org/html/2402.01788v1.
35. Baek, J., Jauhar, S. K., Cucerzan, S. & Hwang, S. J. ResearchAgent: Iterative research idea generation over scientific literature with Large Language Models. *arXiv [cs.CL]* (2024).
36. Si, C., Yang, D. & Hashimoto, T. Can LLMs generate novel research ideas? A large-scale human study with 100+ NLP researchers. *arXiv [cs.CL]* (2024) doi:10.48550/arXiv.2409.04109.
37. Meskó, B. & Görög, M. A short guide for medical professionals in the era of artificial intelligence. *NPJ Digit. Med.* **3**, 126 (2020).
38. Bluemke, D. A. *et al.* Assessing radiology research on artificial intelligence: A brief guide for authors, reviewers, and readers-from the radiology editorial board. *Radiology* **294**, 487–489 (2020).
39. Esteva, A. *et al.* Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **542**, 115–118 (2017).

40. Becker, A. S. *et al.* Classification of breast cancer in ultrasound imaging using a generic deep learning analysis software: a pilot study. *Br. J. Radiol.* **91**, 20170576 (2018).

41. Bien, N. *et al.* Deep-learning-assisted diagnosis for knee magnetic resonance imaging: Development and retrospective validation of MRNet. *PLoS Med.* **15**, e1002699 (2018).

42. Rajpurkar, P. *et al.* Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists. *PLoS Med.* **15**, e1002686 (2018).

43. Fogel, A. L. & Kvedar, J. C. Artificial intelligence powers digital medicine. *NPJ Digit. Med.* **1**, 5 (2018).

44. Ting, D. S. W., Yi, P. H. & Hui, F. Clinical applicability of deep learning system in detecting tuberculosis with chest radiography. *Radiology* **286**, 729–731 (2018).

45. Ting, D. S. W. *et al.* Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. *JAMA* **318**, 2211–2223 (2017).

46. NEI joins call for standardization of ophthalmic imaging devices. https://www.nei.nih.gov/about/news-and-events/news/nei-joins-call-standardization-ophthalmic-imaging-devices (2021).

47. Lee, A. Y. *et al.* Recommendations for Standardization of Images in Ophthalmology. *Ophthalmology* **128**, 969–970 (2021).

48. Dow, E. R. *et al.* From data to deployment: The Collaborative Community on ophthalmic imaging roadmap for artificial intelligence in age-related macular degeneration. *Ophthalmology* **129**, e43–e59 (2022).

49. Baxter, S. L. & Lee, A. Y. Gaps in standards for integrating artificial intelligence technologies into ophthalmic practice. *Curr. Opin. Ophthalmol.* **32**, 431–438 (2021).

50. Radgoudarzi, N. *et al.* Barriers to extracting and harmonizing glaucoma testing data: Gaps, shortcomings, and the pursuit of FAIRness. *Ophthalmol. Sci.* **4**, 100621 (2024).

51. AI-READI Consortium. AI-READI: rethinking AI data collection, preparation and sharing in

diabetes research and beyond. *Nat. Metab.* 1–3 (2024).

52. Gim, N. *et al.* Streamlined DICOM standardization in retinal imaging: Bridging gaps in ophthalmic healthcare and AI research. *Invest. Ophthalmol. Vis. Sci.* **65**, 5879–5879 (2024).
53. Documentation for the AI-READI Dataset. https://docs.aireadi.org/.
54. Artificial Intelligence Ready and Equitable Atlas for Diabetes Insights. *AI-READI* https://aireadi.org/.
55. Pydicom. https://pydicom.github.io/.
56. Zhou, Y. *et al.* A foundation model for generalizable disease detection from retinal images. *Nature* **622**, 156–163 (2023).
57. Liu, Z. *et al.* OCTCube: A 3D foundation model for optical coherence tomography that improves cross-dataset, cross-disease, cross-device and cross-modality analysis. *arXiv [eess.IV]* (2024).
58. Shi, D. *et al.* EyeFound: A multimodal generalist foundation model for ophthalmic imaging. *arXiv [cs.CV]* (2024).
59. Park, W. Y. *et al.* Development of Medical Imaging data standardization for imaging-based observational research: OMOP common data model extension. *J Imaging Inform Med* **37**, 899–908 (2024).
60. Shah, A., Muddana, P. S. & Halabi, S. A review of core concepts of imaging informatics. *Cureus* **14**, e32828 (2022).
61. DICOM vs PACS: Streamlining Healthcare Beyond Differences. *postDICOM* https://www.postdicom.com/en/blog/dicom-vs-pacs.
62. McKinney, W. Data Structures for Statistical Computing in Python. in *Proceedings of the Python in Science Conference* 56–61 (SciPy, 2010).
63. Ale, L., Gentleman, R., Sonmez, T. F., Sarkar, D. & Endres, C. nhanesA: achieving transparency and reproducibility in NHANES research. *Database (Oxford)* **2024**, baae028 (2024).
64. Hub, F. D. I. Documentation for the AI-READI Dataset. https://docs.aireadi.org/.
65. Bidgood, W. D., Jr, Horii, S. C., Prior, F. W. & Van Syckle, D. E. Understanding and using DICOM,

the data interchange standard for biomedical imaging. *J. Am. Med. Inform. Assoc.* **4**, 199–212 (1997).

66. Mason, D. SU-E-T-33: Pydicom: An open source DICOM library. *Med. Phys.* **38**, 3493–3493 (2011).
67. GBD 2019 Blindness and Vision Impairment Collaborators & Vision Loss Expert Group of the Global Burden of Disease Study. Causes of blindness and vision impairment in 2020 and trends over 30 years, and prevalence of avoidable blindness in relation to VISION 2020: the Right to Sight: an analysis for the Global Burden of Disease Study. *Lancet Glob Health* **9**, e144–e160 (2021).
68. Vision Loss Expert Group of the Global Burden of Disease Study & GBD 2019 Blindness and Vision Impairment Collaborators. Global estimates on the number of people blind or visually impaired by cataract: a meta-analysis from 2000 to 2020. *Eye* (2024) doi:10.1038/s41433-024-02961-1.
69. Khairallah, M. *et al.* Number of People Blind or Visually Impaired by Cataract Worldwide and in World Regions, 1990 to 2010. *Invest. Ophthalmol. Vis. Sci.* **56**, 6762–6769 (2015).
70. Rossi, T. *et al.* Cataract surgery practice patterns worldwide: a survey. *BMJ Open Ophthalmol* **6**, e000464 (2021).
71. Chen, X., Xu, J., Chen, X. & Yao, K. Cataract: Advances in surgery and whether surgery remains the only treatment in future. *Adv Ophthalmol Pract Res* **1**, 100008 (2021).
72. Stein, J. D., Khawaja, A. P. & Weizer, J. S. Glaucoma in Adults-Screening, Diagnosis, and Management: A Review. *JAMA* **325**, 164–174 (2021).
73. Sheybani, A. *et al.* Open-angle glaucoma: Burden of illness, current therapies, and the management of nocturnal IOP variation. *Ophthalmol. Ther.* **9**, 1–14 (2020).
74. Vajaranant, T. S., Wu, S., Torres, M. & Varma, R. The changing face of primary open-angle glaucoma in the United States: demographic and geographic changes from 2011 to 2050. *Am. J. Ophthalmol.* **154**, 303–314.e3 (2012).
75. Rhee, D. J. Primary Open-Angle Glaucoma. *Merck Manual Professional Edition*

https://www.merckmanuals.com/professional/eye-disorders/glaucoma/primary-open-angle-glaucoma.

76. Glaucoma: The 'silent thief' begins to tell its secrets. https://www.nei.nih.gov/about/news-and-events/news/glaucoma-silent-thief-begins-tell-its-secrets (2014).

77. Hollands, H. *et al.* Do findings on routine examination identify patients at risk for primary open-angle glaucoma?: The rational clinical examination systematic review. *JAMA* **309**, 2035 (2013).

78. Gim, N. *et al.* Elevated intraocular pressure immediately after cataract surgery and future risk of primary open-angle glaucoma in the IRIS® Registry (Intelligent Research in Sight). *Ophthalmol. Sci.* 100851 (2025).

79. Coban-Karatas, M., Sizmaz, S., Altan-Yaycioglu, R., Canan, H. & Akova, Y. A. Risk factors for intraocular pressure rise following phacoemulsification. *Indian J. Ophthalmol.* **61**, 115–118 (2013).

80. Annam, K. *et al.* Risk Factors for Early Intraocular Pressure Elevation After Cataract Surgery in a Cohort of United States Veterans. *Mil. Med.* **183**, e427–e433 (2018).

81. Ahmed, I. I. K., Kranemann, C., Chipman, M. & Malam, F. Revisiting early postoperative follow-up after phacoemulsification. *J. Cataract Refract. Surg.* **28**, 100–108 (2002).

82. Lidder, A. K., Vanner, E. A., Chang, T. C., Lum, F. & Rothman, A. L. Intraocular Pressure Spike Following Stand-Alone Phacoemulsification in the IRIS® Registry (Intelligent Research in Sight). *Ophthalmology* **131**, 780–789 (2024).

83. Rothman, A. L., Chang, T. C., Lum, F. & Vanner, E. A. Intraocular Pressure Changes Following Stand-Alone Phacoemulsification: An IRISR Registry Analysis. *Am. J. Ophthalmol.* **245**, 25–36 (2023).

84. Grzybowski, A. & Kanclerz, P. Early postoperative intraocular pressure elevation following cataract surgery. *Curr. Opin. Ophthalmol.* **30**, 56–62 (2019).

85. Slabaugh, M. A., Bojikian, K. D., Moore, D. B. & Chen, P. P. Risk factors for acute postoperative intraocular pressure elevation after phacoemulsification in glaucoma patients. *J. Cataract Refract. Surg.* **40**, 538–544 (2014).

86. Tranos, P., Bhar, G. & Little, B. Postoperative intraocular pressure spikes: the need to treat. *Eye* **18**, 673–679 (2004).

87. Fogagnolo, P. *et al.* Short-term changes in intraocular pressure after phacoemulsification in glaucoma patients. *Ophthalmologica* **228**, 154–158 (2012).

88. Levkovitch-Verbin, H. *et al.* Intraocular pressure elevation within the first 24 hours after cataract surgery in patients with glaucoma or exfoliation syndrome. *Ophthalmology* **115**, 104–108 (2008).

89. Leal, I. *et al.* Intraocular Pressure Reduction After Real-world Cataract Surgery. *J. Glaucoma* **29**, 689–693 (2020).

90. Chen, P. P. *et al.* The Effect of Phacoemulsification on Intraocular Pressure in Glaucoma Patients: A Report by the American Academy of Ophthalmology. *Ophthalmology* **122**, 1294–1307 (2015).

91. Shingleton, B. J., Pasternack, J. J., Hung, J. W. & O'Donoghue, M. W. Three and five year changes in intraocular pressures after clear corneal phacoemulsification in open angle glaucoma patients, glaucoma suspects, and normal patients. *J. Glaucoma* **15**, 494–498 (2006).

92. Shrivastava, A. & Singh, K. The effect of cataract extraction on intraocular pressure. *Curr. Opin. Ophthalmol.* **21**, 118–122 (2010).

93. Poley, B. J., Lindstrom, R. L., Samuelson, T. W. & Schulze, R., Jr. Intraocular pressure reduction after phacoemulsification with intraocular lens implantation in glaucomatous and nonglaucomatous eyes: evaluation of a causal relationship between the natural lens and open-angle glaucoma. *J. Cataract Refract. Surg.* **35**, 1946–1955 (2009).

94. Yoo, C., Amoozgar, B., Yang, K.-S., Park, J.-H. & Lin, S. C. Glaucoma severity and intraocular pressure reduction after cataract surgery in eyes with medically controlled glaucoma. *Medicine* **97**, e12881 (2018).

95. IRIS Registry. https://www.aao.org/iris-registry.

96. Chiang, M. F., Sommer, A., Rich, W. L., Lum, F. & Parke, D. W., 2nd. The 2016 American Academy of Ophthalmology IRIS® Registry (Intelligent Research in Sight) Database: Characteristics and Methods. *Ophthalmology* **125**, 1143–1148 (2018).

97. Lee, C. S. *et al.* Smoking Is Associated with Higher Intraocular Pressure Regardless of Glaucoma: A Retrospective Study of 12.5 Million Patients Using the Intelligent Research in Sight (IRIS®) Registry. *Ophthalmol Glaucoma* **3**, 253–261 (2020).

98. Saraf, S. S. *et al.* Demographics and Seasonality of Retinal Detachment, Retinal Breaks, and Posterior Vitreous Detachment from the Intelligent Research in Sight Registry. *Ophthalmol Sci* **2**, 100145 (2022).

99. Lacy, M. *et al.* Endophthalmitis Rate in Immediately Sequential versus Delayed Sequential Bilateral Cataract Surgery within the Intelligent Research in Sight (IRIS®) Registry Data. *Ophthalmology* **129**, 129–138 (2022).

100. Yasutani, H., Hayashi, K., Hayashi, H. & Hayashi, F. Intraocular pressure rise after phacoemulsification surgery in glaucoma patients. *J. Cataract Refract. Surg.* **30**, 1219–1224 (2004).

101. Hayashi, K., Yoshida, M., Manabe, S.-I. & Yoshimura, K. Prophylactic Effect of Oral Acetazolamide against Intraocular Pressure Elevation after Cataract Surgery in Eyes with Glaucoma. *Ophthalmology* **124**, 701–708 (2017).

102. Tranos, P. G. *et al.* Same-day versus first-day review of intraocular pressure after uneventful phacoemulsification. *J. Cataract Refract. Surg.* **29**, 508–512 (2003).

103. Lee, J. Y. *et al.* Aqueous humour outflow imaging: seeing is believing. *EYE* **35**, 202–215 (2021).

104. Malvankar-Mehta, M. S., Fu, A., Subramanian, Y. & Hutnik, C. Impact of Ophthalmic Viscosurgical Devices in Cataract Surgery. *J. Ophthalmol.* **2020**, 7801093 (2020).

105. Rainer, G. *et al.* Natural course of intraocular pressure after cataract surgery with sodium hyaluronate 1% versus hydroxypropylmethylcellulose 2%. *Ophthalmology* **114**, 1089–1093 (2007).

106. Rainer, G. *et al.* Intraocular pressure after small incision cataract surgery with Healon5 and Viscoat. *J. Cataract Refract. Surg.* **26**, 271–276 (2000).

107. Katz, E. A., Majmudar, S. & Aref, A. A. Prophylaxis and treatment of acute intraocular pressure rise after cataract surgery: considerations to aid in decision-making. *Expert Rev. Clin. Pharmacol.* **17**, 995–997 (2024).

108. Jayaram, H., Kolko, M., Friedman, D. S. & Gazzard, G. Glaucoma: now and beyond. *Lancet* **402**, 1788–1801 (2023).

109. Tham, Y.-C. *et al.* Global prevalence of glaucoma and projections of glaucoma burden through 2040: a systematic review and meta-analysis. *Ophthalmology* **121**, 2081–2090 (2014).

110. Kass, M. A. *et al.* The Ocular Hypertension Treatment Study: a randomized trial determines that topical ocular hypotensive medication delays or prevents the onset of primary open-angle glaucoma. *Arch. Ophthalmol.* **120**, 701–13; discussion 829–30 (2002).

111. Goldstein, N. D., Kahal, D., Testa, K., Gracely, E. J. & Burstyn, I. Data Quality in Electronic Health Record Research: An Approach for Validation and Quantitative Bias Analysis for Imperfectly Ascertained Health Outcomes Via Diagnostic Codes. *Harv Data Sci Rev* **4**, (2022).

112. Holmes, J. H. *et al.* Why Is the Electronic Health Record So Challenging for Research and Clinical Care? *Methods Inf. Med.* **60**, 32–48 (2021).

113. Bell, S. K. *et al.* Frequency and Types of Patient-Reported Errors in Electronic Health Record Ambulatory Care Notes. *JAMA Netw Open* **3**, e205867 (2020).

114. Agniel, D., Kohane, I. S. & Weber, G. M. Biases in electronic health record data due to processes within the healthcare system: retrospective observational study. *BMJ* **361**, k1479 (2018).

115. Skuta, G. L., Ding, K., Lum, F. & Coleman, A. L. An IRIS Registry-based assessment of primary open-angle glaucoma practice patterns in academic versus nonacademic settings. *Am. J. Ophthalmol.* **242**, 228–242 (2022).

116. Vu, D. M. *et al.* Risk factors for glaucoma diagnosis and surgical intervention following pediatric cataract surgery in the IRIS® registry. *Ophthalmol. Glaucoma* **7**, 131–138 (2024).

117. CMS. https://www.cms.gov/newsroom/data.

118. Johnson, A. E. W. *et al.* MIMIC-III, a freely accessible critical care database. *Sci Data* **3**, 160035 (2016).

119. Stonebraker, M. & Kemnitz, G. The POSTGRES next generation database management system. *Commun. ACM* **34**, 78–92 (1991).

120. Brown, A. & Wilson, G. *The Architecture of Open Source Applications, Volume II*. (Lulu.com, Barking, England, 2012).

121. Wang, W. *et al.* MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *arXiv [cs.CL]* (2020) doi:10.48550/arXiv.2002.10957.

122. Therneau, T. A package for survival analysis in R. *R package version* **2**, 2014 (2015).

123. Lumley, T. Analysis of complex survey samples. *J. Stat. Softw.* **9**, 1–19 (2004).

124. van Buuren, S. & Groothuis-Oudshoorn, K. mice: Multivariate Imputation by Chained Equations inR. *J. Stat. Softw.* **45**, 1–67 (2011).

125. Guo, W. *et al.* Systemic immune-inflammation index is associated with diabetic kidney disease in Type 2 diabetes mellitus patients: Evidence from NHANES 2011-2018. *Front. Endocrinol. (Lausanne)* **13**, 1071465 (2022).

126. Montgomery, J., Lu, J., Ratliff, S. & Mezuk, B. Food insecurity and depression among adults with diabetes: Results from the National Health and Nutrition Examination Survey (NHANES). *Diabetes Educ.* **43**, 260–271 (2017).

127. Wang, J., Zhou, D. & Li, X. The association between neutrophil-to-lymphocyte ratio and diabetic depression in U.s. adults with diabetes: Findings from the 2009-2016 National Health and Nutrition Examination Survey (NHANES). *Biomed Res. Int.* **2020**, 8297628 (2020).

128. Wang, S. V., Sreedhara, S. K., Schneeweiss, S. & REPEAT Initiative. Reproducibility of real-world evidence studies using clinical practice data to inform regulatory and coverage decisions. *Nat. Commun.* **13**, 5126 (2022).

129. Zhang, X. *et al.* Prevalence of diabetic retinopathy in the United States, 2005-2008. *JAMA* **304**, 649–656 (2010).

130. Wang, H., Guo, Z. & Xu, Y. Association of monocyte-lymphocyte ratio and proliferative diabetic retinopathy in the U.S. population with type 2 diabetes. *J. Transl. Med.* **20**, 219 (2022).

131. Wang, G.-X. *et al.* The correlation between serum albumin and diabetic retinopathy among people with type 2 diabetes mellitus: NHANES 2011-2020. *PLoS One* **17**, e0270019 (2022).

132. Xu, W., Xu, X., Zhang, M. & Sun, C. Association between HDL cholesterol with diabetic retinopathy in diabetic patients: a cross-sectional retrospective study. *BMC Endocr. Disord.* **24**, 65 (2024).

133. Yang, P., Xu, W., Liu, L. & Yang, G. Association of lactate dehydrogenase and diabetic retinopathy in US adults with diabetes mellitus. *J. Diabetes* **16**, e13476 (2024).

134. Li, X. & Chen, M. Correlation of hemoglobin levels with diabetic retinopathy in US adults aged ≥40 years: the NHANES 2005-2008. *Front. Endocrinol. (Lausanne)* **14**, 1195647 (2023).

135. Dai, H., Liu, L. & Xu, W. Association of albumin-to-creatinine ratio with diabetic retinopathy among US adults (NHANES 2009-2016). *Endocrinol. Diabetes Metab.* **8**, e70029 (2025).

136. Mottl, A. K. *et al.* Normoalbuminuric diabetic kidney disease in the U.S. population. *J. Diabetes Complications* **27**, 123–127 (2013).

137. Li, X., Wang, L., Liu, M., Zhou, H. & Xu, H. Association between neutrophil-to-lymphocyte ratio and diabetic kidney disease in type 2 diabetes mellitus patients: a cross-sectional study. *Front. Endocrinol. (Lausanne)* **14**, 1285509 (2023).

138. Wang, Z. *et al.* The relationship between weight-adjusted-waist index and diabetic kidney disease in patients with type 2 diabetes mellitus. *Front. Endocrinol. (Lausanne)* **15**, 1345411 (2024).

139. Liang, Y., Ding, L., Tao, M. & Zhu, Y. The association of metabolic profile of folate with diabetic kidney disease: evidence from 2011-2020 cycles of the NHANES. *Ren. Fail.* **46**, 2420830 (2024).

140. Wang, Y., Lopez, J. M. S., Bolge, S. C., Zhu, V. J. & Stang, P. E. Depression among people with type 2 diabetes mellitus, US National Health and Nutrition Examination Survey (NHANES), 2005-2012. *BMC Psychiatry* **16**, 88 (2016).

141. Casagrande, S. S., Lee, C., Stoeckel, L. E., Menke, A. & Cowie, C. C. Cognitive function among older adults with diabetes and prediabetes, NHANES 2011-2014. *Diabetes Res. Clin. Pract.* **178**, 108939 (2021).

142. Lee, J. *et al.* Prevalence, awareness, treatment, and control of diabetes mellitus by depressive symptom severity: a cross-sectional analysis of NHANES 2011-2016. *BMJ Open Diabetes Res.*

Care **9**, e002268 (2021).

143. Li, B. *et al.* Association of serum uric acid with all-cause and cardiovascular mortality in diabetes. *Diabetes Care* **46**, 425–433 (2023).

144. Ding, L. *et al.* The prognostic value of the stress hyperglycemia ratio for all-cause and cardiovascular mortality in patients with diabetes or prediabetes: insights from NHANES 2005-2018. *Cardiovasc. Diabetol.* **23**, 84 (2024).

145. Al-Kindi, S. G. *et al.* Red cell distribution width is associated with all-cause and cardiovascular mortality in patients with diabetes. *Biomed Res. Int.* **2017**, 5843702 (2017).

146. Song, K., Tan, X., Qin, T., Lu, J. & Liu, T.-Y. MPNet: Masked and permuted pre-training for language understanding. in (2020). doi:10.5555/3495724.3497138.

147. Lim, J. I. *et al.* Diabetic retinopathy preferred practice pattern®. *Ophthalmology* **132**, P75–P162 (2025).

148. Ning, Y. *et al.* Association of prognostic nutritional index with the risk of all-cause mortality and cardiovascular mortality in patients with type 2 diabetes: NHANES 1999-2018. *BMJ Open Diabetes Res. Care* **11**, e003564 (2023).

149. Zhao, P., Du, T., Zhou, Q. & Wang, Y. Association of weight-adjusted-waist index with all-cause and cardiovascular mortality in individuals with diabetes or prediabetes: a cohort study from NHANES 2005-2018. *Sci. Rep.* **14**, 24061 (2024).

150. Haller, C. Hypoalbuminemia in renal failure: pathogenesis and therapeutic considerations. *Kidney Blood Press. Res.* **28**, 307–310 (2005).

151. Cheng, T. *et al.* The level of serum albumin is associated with renal prognosis and renal function decline in patients with chronic kidney disease. *BMC Nephrol.* **24**, 57 (2023).

152. Suchak, T. *et al.* Explosion of formulaic research articles, including inappropriate study designs and false discoveries, based on the NHANES US national health database. *PLoS Biol.* **23**, e3003152 (2025).

153. Leese, P. *et al.* Clinical encounter heterogeneity and methods for resolving in networked EHR data: a

study from N3C and RECOVER programs. *J. Am. Med. Inform. Assoc.* **30**, 1125–1136 (2023).

154. Hur, K. *et al.* Unifying Heterogeneous Electronic Health Records Systems via Text-Based Code Embedding. in *Conference on Health, Inference, and Learning* 183–203 (PMLR, 2022).

155. Voss, E. A. *et al.* Feasibility and utility of applications of the common data model to multiple, disparate observational health databases. *J. Am. Med. Inform. Assoc.* **22**, 553–564 (2015).