Survey-Based Methodologies for Enhanced Assessment of Cause of Death

Shuxian Fan

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

Doctor of Philosophy

University of Washington

2025

Reading Committee:

Tyler H. McCormick, Chair

Emanuela Furfaro

Ema Perkovic

Program Authorized to Offer Degree:
Statistics

University of Washington

Abstract

Survey-Based Methodologies for Enhanced Assessment of Cause of Death

Shuxian Fan

Chair of the Supervisory Committee:
Tyler H. McCormick
Department of Statistics

This dissertation explores several statistical challenges in cause-of-death (COD) assessment from verbal autopsy (VA) surveys—structured interviews with caregivers of the deceased in regions where traditional medical certification is unavailable. Despite their crucial role in mortality surveillance, VA data analysis is complicated by inconsistent age categorization, respondent burden from lengthy questionnaires, and potential biases in automated classification systems.

The first project develops a Bayesian framework for reconciling inconsistent age categories across multiple VA data sources. We formulate age-disaggregated death counts as fully classified multinomial data and show that incorporating partially classified aggregated data can produce an improved Bayes estimator under the Kullback-Leibler (KL) loss. Under specific theoretical conditions, this approach calibrates data with different age structures to generate unified estimates of standardized age distributions. Through numerical studies and applications to real-world mortality data, we demonstrate the method's effectiveness in imputing incomplete classifications and guiding appropriate levels of age disaggregation.

The second project adopts Bayesian active questionnaire design to optimize VA data collection processes. Using posterior-weighted KL information criteria and uncertainty-

aware stopping rules, this approach sequentially selects questions to maximize information while minimizing respondent burden. Validation with gold-standard VA data shows comparable classification accuracy using substantially fewer questions, with implications for improved data collection efficiency.

The third project presents a statistical framework for valid inference using predicted causes from VA narratives. By extending prediction-powered inference (PPI) to multinomial classification, we enable unbiased parameter estimation when using natural language processing models for COD classification. Cross-site validation demonstrates effective correction for transportability errors and highlights the distinction between predictive accuracy and inferential validity.

The last project proposes and validates a proof-of-concept Bayesian mixture model for estimating cause-specific mortality with incomplete age stratification. Using age-mixing proportions within a Bayesian framework, this approach shows that incorporating partially observed age data improves estimation compared to discarding incomplete records. Analysis of demographic survey data from multiple countries reveals that the proposed approach generally yields more accurate cause-specific mortality estimates, with performance advantages varying by the actual age distribution of deaths.

Together, these methodological innovations address fundamental challenges in survey-based mortality surveillance, with applications extending beyond COD assessment to broader problems of inference with incomplete or predicted data.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I extend my deepest gratitude to all those who have provided invaluable support throughout my doctoral journey. First and foremost, I would like to express my sincere appreciation to my advisor, Professor Tyler McCormick. Your guidance in helping me discover the profound connections between statistical methodology and diverse fields of application has profoundly shaped my approach to research. I am deeply grateful to Professor Li Liu and Professor Jamie Perin for your exceptional collaboration and scientific mentorship, and I am privileged to have worked alongside such distinguished researchers. Special recognition is due to Dr. Kentaro Hoffman. Thank you for providing exceptional collaborative support throughout my doctoral studies. Your generous sharing of research insights and methodological expertise has significantly enhanced the quality of this work.

I would also like to thank all the friends I have made at the University of Washington, who have contributed to making my time in Seattle both professionally rewarding and personally fulfilling. The intellectual community fostered within the Department of Statistics has provided an environment conducive to both scholarly growth and meaningful relationships.

Finally, I reserve my most heartfelt gratitude for those closest to me. To my partner, Linquan, whose unwavering support and understanding have sustained me through the most challenging periods of this endeavor. To my parents, Yanxia and Yunli, whose steadfast belief in my abilities and consistent encouragement have provided the foundation upon which this achievement rests. Your collective support has been the bedrock of my perseverance and success.

DEDICATION

To my parents, Yunli and Yanxia, and my best friend and partner, Linquan.

Chapter 1
INTRODUCTION

This dissertation explores several statistical challenges in cause-of-death (COD) assessment from verbal autopsy (VA), which involve surveys—structured interviews with family members or caregivers of deceased individuals in regions lacking formal medical death certification. While these surveys provide essential data for tracking mortality patterns, they present significant methodological challenges, including inconsistent age categorization, respondent burden from lengthy questionnaires, and systematic biases in automated classification systems.

Chapter 2 presents a Bayesian approach to address the problem of mismatched age categories when combining VA data from different sources. The method treats age-specific death counts as outcomes from a multinomial distribution and demonstrates that including partially classified data from aggregated age groups can produce an improved Bayes estimator under the Kullback-Leibler (KL) loss. The framework addresses two scenarios: situations where age categories from one data source are entirely nested within broader categories from another source, and more complex cases where age groupings overlap in non-hierarchical ways. The approach leverages individual-level death registration records, alongside grouped mortality counts, to estimate standardized age distributions across different data sources. Validation through simulation studies and real mortality datasets demonstrates that the method successfully fills gaps in incomplete age classifications, while offering practical recommendations for optimal age group granularity in mortality surveillance systems.

Chapter 3 introduces a Bayesian active questionnaire design to streamline VA questionnaires by intelligently selecting which questions to ask based on responses already

provided. Rather than administering identical, lengthy surveys to all respondents, this framework dynamically selects the most informative next question for each specific case. It terminates the interview when sufficient information has been gathered for a reliable determination of COD. The method personalizes the questioning process to optimize diagnostic accuracy for individual CODs, rather than applying blanket questionnaire reductions across all cases. This adaptive strategy integrates seamlessly with existing probabilistic COD classification methods, working as a preprocessing step regardless of the underlying statistical model used to link symptoms with potential causes. Testing on validated VA datasets demonstrates that the approach maintains comparable diagnostic accuracy while requiring significantly fewer questions per interview, offering substantial improvements in data collection efficiency and reduced burden on grieving families.

Chapter 4 develops a statistical framework that ensures valid inference using predicted causes from VA narratives. VA serves as a critical epidemiological tool for mortality surveillance in settings where complete medical records are unavailable, with COD assignment typically performed either through physician review or automated algorithms that analyze reported symptoms. This chapter extends prediction-powered inference (PPI) methods to handle multinomial classification problems, enabling researchers to obtain unbiased estimates of cause-specific mortality rates even when automated classifiers contain systematic prediction errors. The framework addresses the crucial distinction between a model's ability to classify individual cases correctly and its capacity to conduct valid population-level inference. Validation across different geographic sites demonstrates that the method successfully corrects for biases that arise when applying models trained in one population to data from another, highlighting the distinction between predictive accuracy and inferential validity as fundamentally different aspects of model evaluation in public health surveillance.

Chapter 5 introduces a Bayesian mixture modeling approach to estimate COD in under-five children when age information is incomplete or imprecisely recorded. Accurate assessment of under-five mortality requires detailed age stratification, as risk factors and

disease patterns vary significantly across different developmental periods within the first five years of life. The proposed framework uses mixture models to extract valuable information from cases with partial or uncertain age data rather than discarding them entirely, as is standard practice in traditional complete-case analyses. This approach recognizes that even imprecise age information, such as knowing a child was "less than one year old" without exact age in months, contains useful epidemiological signals that can improve overall mortality estimates. Empirical testing across various datasets and epidemiological scenarios provides strong evidence that incorporating partially observed age data through this mixture modeling framework yields more accurate and robust COD estimates compared to methods that exclude cases with incomplete age information. This demonstrates significant improvements across multiple performance measures and real-world mortality surveillance contexts.

Chapter 6 offers concluding remarks and discusses potential avenues for future research.

Chapter 2

BAYESIAN AGE CATEGORY RECONCILIATION FOR AGE- AND CAUSE-SPECIFIC UNDER-FIVE MORTALITY ESTIMATION

2.1 Introduction

Understanding age- and cause-specific under-five mortality (ACSU5M) is crucial for informing and evaluating age and disease-targeted interventions and policies aimed at reducing child mortality. ACSU5M data are typically collected through demographic and health surveys (DHS), sample registration systems (SRS), and vital registration (VR) systems. A persistent challenge for accurate and reliable ACSU5M estimation is the lack of complete and consistent data, particularly in low-income countries, where VR records are often incomplete or unavailable.

Exacerbating these challenges, ACSU5M estimates mandate age resolution well beyond what's required in adults to effectively target policies and programs. Whereas adults are at risk for some causes (e.g., heart disease) for decades, congenital disabilities are common among neonates but not older children. In most low-resource settings where deaths happen outside of hospitals or are not systematically recorded, researchers and health officials rely on multiple smaller-scale data sources to piece together a comprehensive picture of ACSU5M.

Data are often collected by different organizations and inconsistently aggregated across ages, making it difficult in many settings to perform comprehensive COD analyses across ages (Diaz et al., 2021). Combining data in such a setting would mean discarding any data sources that do not have exactly overlapping age categories, which would be suboptimal both statistically and practically (Fienberg, 1972; Chen and Fienberg,

1976; Williamson and Haber, 1994). Furthermore, when deciding how to aggregate ages, researchers often resort to past studies to choose age categories without considering empirical evidence or matching age categories in other studies. This can result in the use of age groupings that may not accurately reflect the actual age distribution within each COD.

<table>
<tr><td colspan="4">Under-five Child Age Categories</td></tr>
<tr><td rowspan="6">$0-59$ months</td><td rowspan="2">$0-27$ days</td><td rowspan="2">$0-27$ days</td><td>$0-6$ days</td></tr>
<tr><td>$7-27$ days</td></tr>
<tr><td rowspan="4">$1-59$ months</td><td rowspan="2">$1-11$ months</td><td>$1-5$ months</td></tr>
<tr><td>$6-11$ months</td></tr>
<tr><td rowspan="2">$12-59$ months</td><td>$12-23$ months</td></tr>
<tr><td>$24-59$ months</td></tr>
</table>

Table 2.1: Sample age groups from ACSU5M research.

In this work, we introduce a Bayesian approach to calibrating data from multiple sources, which are reported at different levels of disaggregation. We consider both cases where data sources have age categories nested within those present in other sources and cases with non-nested age structures. Our method estimates standardized age group distributions by combining individual registration data, when available, with fully classified age-disaggregated death counts as group counts from a multinomial distribution. We then incorporate partially classified aggregated data to estimate the multinomial probabilities jointly. This approach potentially yields more accurate estimates of age group distributions and age- and cause-specific death counts at the desired level of granularity.

The proposed approach addresses the challenge of concurrently estimating multinomial cell probabilities using both fully and partially classified multinomial samples. Tabulated data can be partially classified by rows or columns, which commonly occurs

when observing incomplete contingency tables (Albert, 1985; Chen and Fienberg, 1976; Gibbons and Greenberg, 1994). While previous work typically focused on partially classified data where the partial classification is limited to row or column sums, we address a more general problem of partially categorized data—where reported age groups represent unions of more refined, fully classified age groups—in sampling situations with categorical observations. This problem arises in various contexts, such as signal detection (Blumenthal, 1968), where observers detect the presence of a signal but cannot distinguish among several possible signal types. Our approach allows partial classifications to be any subsets of the fully classified group index set, enabling us to handle non-nested age structures in the ACSU5M context. We provide a comprehensive Bayesian framework for conducting inference and demonstrate its performance through numerical studies.

Our work extends existing multi-sample approaches (Hocking and Oxspring, 1971, 1974; Oxspring and LaMotte, 1977; Koch et al., 1972) for the simultaneous estimation of multinomial probabilities with completely and partially classified data to more complex cell grouping patterns. While previous approaches have been limited in scope, we develop a framework that accommodates diverse classification structures encountered in real-world mortality data. From a theoretical perspective, we provide decision-theoretic support through numerical studies, demonstrating how integrating partially classified data enhances Bayesian estimators of multinomial probabilities under conditions we explicitly define. This theoretical foundation establishes the conditions under which our proposed estimators are superior to those that discard partially classified data. To validate our methodology in practical settings, we conduct extensive simulation studies based on observed, disaggregated data to assess the performance of our age reconciliation method in ACSU5M studies. These simulations, utilizing realistic data patterns from multiple countries and periods, confirm that our approach provides valuable solutions for mitigating age inconsistencies in ACSU5M studies and enhances the accuracy of cause-specific mortality estimates.

2.2 Related Work

2.2.1 Foundational Approaches to Incomplete Multinomial Data

The problem of incomplete contingency tables and partially classified multinomial data poses a significant challenge in statistical modeling, particularly in contexts such as ACSU5M studies, where age-disaggregation inconsistencies are typical. This section examines the theoretical foundations and recent advances in methodologies for addressing such challenges.

Statistical approaches to incomplete contingency tables broadly fall into two categories: single-sample methods (Blumenthal, 1968; Chen and Fienberg, 1974; Woolson and Clarke, 1984) and multi-sample methods (Hocking and Oxspring, 1971, 1974; Oxspring and LaMotte, 1977; Koch et al., 1972). Single-sample approaches conceptualize fully and partially classified data as originating from a single multinomial distribution, whereas multi-sample methods treat these data types as independent multinomial samples. This fundamental distinction guides the selection of methodology based on data collection characteristics—multi-sample approaches are typically preferred when data incompleteness results from design constraints rather than random missingness. The seminal work by Blumenthal (1968) established maximum-likelihood estimation procedures for incomplete multinomial data where each category shares data with at most one aggregated category. This framework proved particularly valuable for signal detection problems, where observers could detect signals but were unable to distinguish between certain signal types. Building upon this foundation, Hocking and Oxspring (1971) and Hocking and Oxspring (1974) reconceptualized the problem as a complete cross-classified multinomial sample supplemented with margin tables, where cell probabilities represent sums of probabilities from a "parent" multinomial distribution. For example, consider a parent population comprising three categories with probabilities p_1, p_2, p_3, complemented by a supplementary sample with probabilities $p_1 + p_2$ and p_3.The analytical challenge emerges because observed data exist only for individual probabilities p_1, p_2, p_3, and the aggre-

gation $p_1 + p_2$, with no direct observations for other potential aggregations. To address this challenge, Hocking and Oxspring (1974) developed an iterative algorithm for maximum likelihood estimation with incomplete multinomial data. Subsequently, Oxspring and LaMotte (1977) demonstrated that optimal allocation of sample sizes to different categorization levels in nested structures could be formulated as a convex programming problem, significantly advancing the practical application of these methods.

2.2.2 Dimensional Extensions and Parametric Constraints

While initial work focused primarily on one-dimensional classification problems, several researchers expanded these frameworks to higher dimensions. Chen and Fienberg (1974, 1976) investigated restricted parameterizations for two-dimensional contingency tables, and Williamson and Haber (1994) further extended this approach to three-dimensional data structures. These dimensional extensions have particular relevance for ACSU5M studies, where data may be cross-classified by multiple factors beyond age categories, such as geographic region, socioeconomic status, and temporal periods.

For multi-dimensional contingency tables with incompletely cross-classified data, Fuchs (1982) adopted the "marginal conformity" concept introduced by Koch et al. (1972) to develop log-linear analytical approaches. Meanwhile, Koch et al. (1972) and Woolson and Clarke (1984) employed weighted least squares analysis with multi-sample and single-sample methods, respectively, providing alternative estimation techniques that remain valuable in specific analytical contexts. The theoretical underpinnings of maximum likelihood estimation for incomplete data from exponential families were formalized by Sundberg (1974), while Rubin (1974) established a generalized framework for parameter estimation in incomplete-data problems. This generalized framework has proven particularly influential in the development of modern approaches to missing data, including multiple imputation techniques that have gained prominence in recent decades.

2.2.3 Distributional and Bayesian Approaches

Beyond maximum likelihood approaches, distributional methods offer an alternative paradigm for addressing incomplete classification. Turnbull (1976) investigated nonparametric estimation with a distributional formulation for arbitrarily grouped, censored, and truncated data. More recently, Hankin (2010) introduced the Hyper-Dirichlet distribution—a high-dimensional conjugate distribution with the simplex as its sample space—expanding the analytical toolkit for complex multinomial data. The connection between contingency table analysis and ranking problems provides another perspective, as both can be viewed as estimation challenges involving latent probability vectors. The Bradley-Terry model for pairwise comparisons (Bradley and Terry, 1952) exemplifies this connection. Hastie and Tibshirani (1997) extended this model to general classification problems, establishing a unified framework relevant to both ranking and contingency table analysis.

In the Bayesian paradigm, numerous approaches have been developed for categorical data with various forms of censoring or missingness. For 2×2 contingency tables with missing row or column information, substantial Bayesian literature exists (Karson and Wrobleski, 1970; Antelman, 1972; Kaufman and King, 1973; Albert and Gupta, 1983; Gunel, 1984; Smith and Gunel, 1984; Smith et al., 1985; Albert, 1985; Kadane, 1985; Gibbons and Greenberg, 1989, 1994). These approaches vary in prior specification, computational methodology, and assumptions about the missing data mechanism. Dickey et al. (1987) explored nested distribution families generalizing the Dirichlet distribution for multinomial settings with noninformative censoring. For scenarios with informative censoring, methodological contributions have been made by Basu and Pereira (1982), Daniel Mimoso Paulino and Alberto De Bragança Pereira (1992), Paulino and de Braganca Pereira (1995), Walker (1996), Tian et al. (2003), and Jiang and Dickey (2008), among others. Notably, Jiang and Dickey (2008) provided Bayesian frameworks for scenarios beyond the scope of previous methods, attempting to relax restrictive assumptions,

such as the requirement of truthful reporting.

2.2.4 Recent Methodological Advances

Contemporary methodological developments have substantially expanded the toolkit for analyzing incomplete multinomial data, particularly in multilevel contexts relevant to ACSU5M studies. Recent innovations in Bayesian inference for incomplete multinomial data include hierarchical approaches that account for complex missingness patterns while leveraging advances in computational Bayesian statistics (Schmid et al., 2014). The integration of Bayesian methods with multilevel modeling frameworks has proven particularly valuable for addressing the hierarchical nature of mortality data, where geographic regions, periods, or demographic groups naturally cluster observations. Specialized Bayesian information criteria have been developed for incomplete data scenarios, where the penalty term is calibrated based on the actual amount of observed information rather than assuming complete-data sample sizes (Zhao et al., 2024). This adaptation significantly improves model selection in datasets with substantial missingness.

The application of imprecise probability models represents another methodological frontier. The Imprecise Dirichlet Model (IDM) has been extended to multilevel settings, addressing challenges in reliability analysis and systems with hierarchical structures where traditional precise probability models may oversimplify uncertainty quantification. This approach is particularly valuable when prior knowledge is limited or ambiguous (Walley, 1996). In the realm of network meta-analysis, Bayesian multinomial models have been extended to handle unordered categorical outcomes with partially observed data, properly accounting for correlations among counts in mutually exclusive categories while enabling meaningful comparisons across multiple treatments and outcome categories (Schmid et al., 2014). These models have found application in healthcare research, where mortality data from different sources often exhibit varying levels of completeness. For spatiotemporal applications, Bayesian methods for incomplete multinomial data have been developed to assess variation in pathogen-variant diversity, addressing challenges

including low cell counts, incomplete classification due to laboratory issues, and unseen variants (Ahn et al., 2007). These methods incorporate Shannon entropy as a measure of diversity and employ Markov chain Monte Carlo (MCMC) techniques for posterior simulation, demonstrating the versatility of Bayesian approaches to handling incomplete multinomial data in epidemiological contexts.

2.2.5 Relationship to Our Work

Our research bridges two critical methodological streams: the multi-sample approach established by Hocking and Oxspring (1971, 1974); Oxspring and LaMotte (1977) and contemporary Bayesian frameworks for incomplete multinomial data. While previous multi-sample methods focused primarily on nested structures with disjoint unions, our work extends this approach to non-nested age structures commonly encountered in ACSU5M studies, where age categories from different data sources may overlap in complex patterns.

A distinguishing feature of our contribution is the development of a comprehensive Bayesian framework that provides both theoretical foundations and practical implementation strategies for estimating multinomial probabilities from datasets with diverse age categorizations. Through decision-theoretic analysis, we demonstrate that incorporating partially classified data can yield superior estimators compared to approaches that discard such information, establishing explicit conditions under which this improvement occurs. Our simulation studies based on realistic mortality data patterns validate the practical utility of these methods in ACSU5M contexts, offering insights into the impact of different age-disaggregation patterns on estimation accuracy.

This work addresses a critical gap in mortality research methodology, where inconsistent age categorizations across data sources have historically complicated efforts to develop comprehensive, temporally and geographically robust mortality estimates. Unlike previous approaches, which often require specialized computational techniques for each application context, our Bayesian framework provides a flexible and generalizable methodology applicable across diverse data structures and missingness patterns. This

flexibility, combined with the theoretical guarantees established through decision theory, positions our approach as a valuable contribution to both the methodological literature on incomplete multinomial data and practical applications in global health and demographic research.

2.3 Method

2.3.1 Problem Statement

Let $\mathbf{X} = (X_i)_{i=1}^k$ be the fully classified observations that follow a multinomial distribution with parameters $(N, \boldsymbol{\theta})$,

$$\mathbf{X} \sim \text{Multi}(\mathbf{x}|N, \boldsymbol{\theta}) = \frac{N!}{\prod_{i=1}^k x_i!} \prod_{i=1}^k \theta_i^{x_i},$$

where $\boldsymbol{\theta} = \{\theta_i\}_{i=1}^k$ is the unknown vector of group probabilities and $\mathbf{x} = (x_i)_{i=1}^k \in \{(\mathbf{x}')_{i=1}^k \in \mathbb{N}_0 | \sum_{i=1}^k x_i' = N\}$, with $\mathbb{N}_0 = \{0\} \cup \mathbb{N} = \{0, 1, 2, \dots\}$. In ACSU5M studies, these could represent age- and cause-specific death counts observed at a desired disaggregated level with k different groups (Diaz et al., 2021).

We also consider additional data that are partially classified by age categories. That is, $\mathbf{Y}' = (Y_i')_{i=1}^k \sim \text{Multi}(N', \boldsymbol{\theta})$ independently of $\mathbf{X}$, but we only observe the partially classified data $\mathbf{Y} = (Y_{A_j})_{j=0}^m$ that follow a multinomial distribution with parameters $(N', \boldsymbol{\tau})$, where $(A_j)_{j=1}^m$ are the distinct non-singleton proper subsets of $S = \{1, \dots, k\}$, $A_0 = S - \cup_{j=1}^m A_j$, and $\boldsymbol{\tau} = (\tau_j)_{j=0}^m$ with $\tau_j = \sum_{i \in A_j} \theta_i$. Note that A_0 can be the empty set $\emptyset$.

The problem of simultaneously estimating multinomial probabilities with partially classified data has been studied in various contexts (Fienberg and Holland, 1973; Leonard, 1977; Alam and Mitra, 1986; Albert, 1987). However, our approach extends this literature by addressing both nested and non-nested partial classifications, which are common in public health data collection systems (Clark et al., 2018b; Fan et al., 2023).

To illustrate our notation in the context of age disaggregation for child mortality

data, consider the following standardized age categories commonly used in global health monitoring (Schumacher et al., 2020):

$$\begin{cases} \{0-27 \text{ days}\} := \{1,2\} & \begin{cases} \{0-6 \text{ days}\} := \{1\} \\ \{7-27 \text{ days}\} := \{2\} \end{cases} \\ \{1-11 \text{ months}\} := \{3,4\} & \begin{cases} \{1-5 \text{ months}\} := \{3\} \\ \{6-11 \text{ months}\} := \{4\} \end{cases} \\ \{12-59 \text{ months}\} := \{5,6\} & \begin{cases} \{12-23 \text{ months}\} := \{5\} \\ \{24-59 \text{ months}\} := \{6\} \end{cases} \end{cases}.$$

From this classification structure, we define:

$$S = \{1,2,3,4,5,6\},$$
$$A_1 = \{1,2\}, A_2 = \{3,4\}, A_3 = \{5,6\}.$$

This hierarchical structure represents a nested age classification commonly found in DHS and other mortality data sources. The fully disaggregated age groups (e.g., $0-6$ days, $7-27$ days) represent the finest level of classification (S), while the aggregated groups (e.g., $0-27$ days, $1-11$ months) represent partial classifications (A_j).

We address the problem of estimating $\boldsymbol{\theta}$ with the prior distribution for $\boldsymbol{\theta}$ being the Dirichlet distribution with parameter $\boldsymbol{\alpha} = (\alpha_i)_{i=1}^k$ under the KL loss, which can be directly interpreted as a divergence measure induced by entropy (Nayak and Naik, 1989):

$$\mathcal{L}(\mathbf{p}, \boldsymbol{\theta}) = \sum_{i=1}^{k} \theta_i \log \frac{\theta_i}{p_i},$$

where $\mathbf{p} = (p_i)_{i=1}^k \in \Theta = \{(\theta_i)_{i=1}^k \in (0,\infty)^k | \sum_{i=1}^k \theta_i = 1\}$.

The problem is further decomposed into two distinct scenarios, each requiring different analytical approaches: (i)where the sets A_j are mutually exclusive (i.e., form a partition

of S); and (ii) where the sets A_j may overlap, representing more complex classification structures.

The mutually exclusive case corresponds to nested hierarchical classifications commonly found in standardized reporting systems. In contrast, the overlapping case addresses more complex real-world data integration challenges where different data sources may use incompatible classification schemes (Fan et al., 2023; Mulick et al., 2021). Both scenarios present unique statistical challenges, particularly in determining whether incorporating partially classified data leads to improved estimation compared to discarding such incomplete observations.

2.3.2 Bayes Estimator with Disjoint Partial Classifications

Let $\boldsymbol{\rho}_{A_j} = (\theta_i/\tau_j, i \in A_j)^T$, $j = 0, \ldots, m$ be the conditional probabilities of each cell within each group A_j, and let $\boldsymbol{\eta} = (\boldsymbol{\tau}, \boldsymbol{\rho}_{A_j}, j = 0, \ldots, m)$ represent an alternative parameterization of our model. The vector $\boldsymbol{\tau} = (\tau_j)_{j=0}^m$ contains the marginal probabilities for each group A_j, while $\boldsymbol{\rho}_{A_j}$ specifies the conditional distribution within each group. To formulate the likelihood function, we denote $(\boldsymbol{X}, \boldsymbol{Y}) = (X_1, X_2, \ldots, X_k, Y_{A_1}, \ldots, Y_{A_m})$ as our observed data. The probability mass function of $(\boldsymbol{X}, \boldsymbol{Y}) = (\boldsymbol{x}, \boldsymbol{y})$ is given by:

$$f(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{\eta}) = \frac{(N!)(N'!)}{(\prod_{i=1}^k x_i!)(\prod_{j=0}^m y_{A_j}!)} \prod_{j=0}^m \tau_j^{x_{A_j}+y_{A_j}} \prod_{j=0}^m \prod_{r \in A_j} (\rho_{A_j}^{(r)})^{x_r},$$

where $x_{A_j} = \sum_{r \in A_j} x_r$ represents the sum of the fully observed counts in group A_j.

This likelihood formulation extends the standard multinomial likelihood to incorporate both fully classified observations (represented by $\boldsymbol{X}$) and partially classified observations (represented by $\boldsymbol{Y}$), while maintaining the constraints imposed by the group structure. We consider a conjugate Dirichlet prior distribution for $\boldsymbol{\theta}$ with hyperparameter vector $\boldsymbol{\alpha}$:

$$f(\boldsymbol{\theta}|\boldsymbol{\alpha}) \propto \prod_{i=1}^k \theta_i^{\alpha_i - 1},$$

where $\alpha_i > 0$ for all $i \in \{1, \ldots, k\}$. The Dirichlet prior is a natural choice for multinomial probabilities as it allows for flexible modeling of prior beliefs about the distribution of $\boldsymbol{\theta}$ while maintaining mathematical tractability (Teh et al., 2010).

Through application of Bayes' rule and algebraic manipulation, we can derive the posterior density function for the reparameterized model:

$$f(\boldsymbol{\eta}|\boldsymbol{x}, \boldsymbol{y}) \propto \tau_0^{x_{A_0}+\alpha_{A_0}-1} \prod_{j=1}^{m} \tau_j^{x_{A_j}+y_{A_j}+\alpha_{A_j}-1} \times \prod_{j=0}^{m} \prod_{r \in A_j} \left(\rho_{A_j}^{(r)}\right)^{x_r+\alpha_r-1},$$

where $\alpha_{A_j} = \sum_{r \in A_j} \alpha_r$ represents the sum of the prior hyperparameters corresponding to group A_j.

This posterior density exhibits a handy form for statistical inference. Since $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ have a one-to-one relationship, both represent equivalent parameterizations of our model. From Equation (2.1), we can observe that $\boldsymbol{\tau}$ and $(\boldsymbol{\rho}_{A_j})_{j=0}^{m}$ are jointly independent in the posterior distribution. Specifically, for each $j \in \{0, 1, \ldots, m\}$, the conditional probability vector $(\rho_{A_j}^{(r)})_{r \in A_j = \{j_1, \ldots, j_{n_j}\}}$ follows a Dirichlet distribution with parameter vector $(\alpha_{j_1} + x_{j_1}, \ldots, \alpha_{j_{n_j}} + x_{j_{n_j}})$. This result extends the well-known conjugacy property of the Dirichlet-multinomial model to our partially observed data setting (Leonard, 1977; Albert, 1985). Moreover, the marginal probabilities $\boldsymbol{\tau}$ follow a Dirichlet distribution with parameter vector $(x_{A_0} + \alpha_{A_0}, x_{A_1} + y_{A_1} + \alpha_{A_1}, \ldots, x_{A_m} + y_{A_m} + \alpha_{A_m})$. This elegant decomposition of the posterior distribution facilitates efficient computation of posterior quantities of interest.

Given the posterior distribution derived above, we can now determine the Bayes estimator for $\boldsymbol{\theta}$ under the KL loss function. The following lemma provides the closed-form expression for this estimator.

Lemma 1. *With respect to the Dirichlet prior with parameter $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_k)$, given data consist of fully-classified $\boldsymbol{x} = (x_i)_{i=1}^{k}$ and partially-classified $\boldsymbol{y} = (y_{A_j})_{j=0}^{m}$ with dis-*

joint A_j's, the Bayes estimator $(\hat{\theta}_i)_{i=1}^k$, for $i \in A_j$, is given by

$$\hat{\theta}_i = \frac{\alpha_i + x_i}{x_{A_j} + \alpha_{A_j}} \frac{y_{A_j} + \alpha_{A_j} + x_{A_j}}{N + N' + \alpha_S},$$

where $\alpha_S = \sum_{i=1}^k \alpha_i$ represents the sum of all prior hyperparameters.

This estimator has an intuitive interpretation: the first fraction represents the conditional probability of category i within group A_j, which is informed by the fully observed data $\boldsymbol{x}$ and the prior $\boldsymbol{\alpha}$. The second fraction represents the marginal probability of group A_j, which incorporates information from both the fully observed data $\boldsymbol{x}$ and the partially observed data $\boldsymbol{y}$, as well as the prior. The structure of this estimator reveals how information from the partially classified observations contributes to the estimation of $\boldsymbol{\theta}$. When y_{A_j} is large relative to x_{A_j}, the partially classified data substantially influences the estimated probability mass assigned to group A_j. Conversely, within each group A_j, the distribution of probability mass among individual categories $i \in A_j$ is determined solely by the fully observed data and the prior.

This result generalizes previous work on Bayesian estimation with incomplete categorical data (Chen and Fienberg, 1976; Albert, 1985) to a more flexible framework that accommodates arbitrary grouping structures while maintaining computational tractability. The estimator can be directly applied to the age disaggregation problem in demographic and health studies, where different data sources may report deaths at varying levels of age granularity (Fan et al., 2023; Schumacher et al., 2020).

2.3.3 Decision-theoretic Justification for Age Reconciliation

To provide a rigorous decision-theoretic rationale for incorporating partially-classified data in multinomial probability estimation, we conduct a comparative analysis of risk functions. We compare two estimators: $\hat{\boldsymbol{\theta}}$, which incorporates both fully and partially classified data as derived in Lemma 1, and $\tilde{\boldsymbol{\theta}}$, which uses only the fully-classified data $\boldsymbol{x}$.

The Bayes estimator $\tilde{\boldsymbol{\theta}}$ is given by:

$$\tilde{\theta}_i = \frac{x_i + \alpha_i}{\alpha_S + N},$$

which follows directly from the conjugate property of the Dirichlet prior for multinomial distributions (Bernardo and Smith, 2009). This standard form represents the posterior mean when only fully classified observations are available.

To evaluate the relative performance of these estimators, we define the risk difference:

$$\Delta_{\boldsymbol{\theta}}(N, N') = \mathbb{E}_{\boldsymbol{\theta}}[L(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta})] - \mathbb{E}_{\boldsymbol{\theta}}[L(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta})],$$

where the expectation is taken concerning the true parameter $\boldsymbol{\theta}$. A negative value of $\Delta_{\boldsymbol{\theta}}(N, N')$ indicates that $\hat{\boldsymbol{\theta}}$ has lower risk than $\tilde{\boldsymbol{\theta}}$, suggesting that incorporating partially-classified data improves estimation accuracy.

The following lemma provides an insightful decomposition of the risk difference.

Lemma 2. *The risk difference between estimators $\hat{\boldsymbol{\theta}}$ and $\tilde{\boldsymbol{\theta}}$ can be decomposed as:*

$$\begin{aligned}\Delta_{\boldsymbol{\theta}}(N, N') &= \mathbb{E}_{\boldsymbol{\theta}}\left[\log\frac{N' + N + \alpha_S}{N + \alpha_S} + \sum_{j=0}^{m} \theta_{A_j} \log\left\{\frac{x_{A_j} + \alpha_{A_j}}{y_{A_j} + x_{A_j} + \alpha_{A_j}}\right\}\right] \\ &= \sum_{u=1}^{N'} \Delta_{\boldsymbol{\theta}}(N + u - 1, 1).\end{aligned}$$

This decomposition reveals that the total risk difference when including N' partially-classified observations can be expressed as the sum of incremental risk differences from adding each observation individually. The first term in the expectation, $\log\frac{N'+N+\alpha_S}{N+\alpha_S}$, is always positive and increases with N', reflecting the increased uncertainty from incorporating additional partially-classified data. The second term captures the information gained from these observations and is typically negative when the partially-classified data provides valuable information about the group probabilities θ_{A_j}.

Without loss of generality, we focus on analyzing $\Delta_{\boldsymbol{\theta}}(N, 1)$, which represents the risk difference when adding a single partially-classified observation. The following lemma identifies the parameter configuration that maximizes this risk difference.

Lemma 3. *Suppose that* $\min_j \alpha_{A_j} \geq 2$*, then the risk difference* $\Delta_{\boldsymbol{\theta}}(N,1)$ *is maximized at* $\boldsymbol{\theta} = \boldsymbol{\theta}^*$*:*

$$\max_{\boldsymbol{\theta}\in\Theta} \Delta_{\boldsymbol{\theta}}(N,1) = \Delta_{\boldsymbol{\theta}^*}(N,1),$$

where $\boldsymbol{\theta}^* = (\theta_i^*)_{i=0}^k \in \boldsymbol{\Theta}$*,* $\theta^*_{i:i\in A_j} = \frac{1}{(1+m)n_j}$ *for* $i \in A_j = \{j_1, \ldots, j_{n_j}\}$*, and* $\sum_{i=0}^k \theta^*_{i:i\in A_j} = \theta_{A_j} = \frac{1}{m+1}$ *for all* $j = 0, 1, \ldots, m$*.*

This maximum risk configuration $\boldsymbol{\theta}^*$ represents a uniform distribution of probability mass across groups (each group receives equal probability $1/(m+1)$) and within each group (each element within a group receives equal probability). This "maximally uniform" distribution represents the most challenging scenario for reconciling partially-classified data, as it maximizes the entropy of the distribution (Kullback, 1997; Cover, 1999).

Building upon the previous lemmas, we establish the conditions under which $\hat{\boldsymbol{\theta}}$ dominates $\tilde{\boldsymbol{\theta}}$, meaning that $\hat{\boldsymbol{\theta}}$ has lower risk for all possible values of $\boldsymbol{\theta}$.

Theorem 1. *(i) Fix* $m \in \mathbb{N}$ *and* $N \in \mathbb{N}$*. Suppose we have* A_j*'s and* α_i*'s such that* $\min_j \alpha_{A_j} \geq 2$*, then* $\hat{\boldsymbol{\theta}}$ *dominates* $\tilde{\boldsymbol{\theta}}$ *when* N' *is sufficiently large.*

(ii) Fix $m \in \mathbb{N}$ *and* $N' \in \mathbb{N}$*. Suppose we have* A_j*'s and* α_i*'s such that* $\min_j \alpha_{A_j} \geq 2$*, then* $\hat{\boldsymbol{\theta}}$ *dominates* $\tilde{\boldsymbol{\theta}}$ *when* N *is sufficiently large.*

This theorem provides the sufficient condition for $\tilde{\boldsymbol{\theta}}$ to be dominated by $\hat{\boldsymbol{\theta}}$, requiring that $\alpha_{A_j} = \sum_{i\in A_j} \alpha_i \geq 2$ for all $j = 0, \ldots, m$. The condition ensures that there is sufficient prior information about each group to incorporate the partially-classified data properly.

The theoretical results have important practical implications for ACSU5M studies. For non-informative priors, the condition for dominance translates to structural requirements on the classification groups. Specifically, with a uniform prior, we need $\min_j |A_j| \geq 2$, while Jeffrey's prior requires $\min_j |A_j| \geq 4$. With m fixed (the number of distinct groups), these conditions constrain the minimum number of classes k to

be at least $2(m+1)$ with the uniform prior and $4(m+1)$ with Jeffrey's prior, since $\sum_{j=0}^{m} |A_j| = k$. These requirements ensure sufficient granularity in the age categorization to overcome uncertainties introduced by partial classifications.

While these constraints might initially appear restrictive, they align well with the practical structure of ACSU5M data. In typical ACSU5M studies, the observation age range is fixed between 0 and 5 years, and partially classified age groups are typically reported in standardized aggregated categories (Diaz et al., 2021; Schumacher et al., 2020). The required granularity ensures that meaningful information can be extracted from partially classified data, thereby improving estimation. The value of k can also be interpreted as a truncation level in nonparametric Bayesian modeling. This connection relates to established results on truncation bounds for infinite-dimensional models such as the Dirichlet Process (Ishwaran and James, 2001) and the Indian Buffet Process (Doshi et al., 2009), where the bound decreases with the truncation level. This perspective provides additional theoretical justification for our approach within the broader Bayesian nonparametric literature.

The Dirichlet parameter vector $\boldsymbol{\alpha}$ plays a crucial role in our framework, capturing prior beliefs about $\boldsymbol{\theta}$. Following the interpretation of Teh et al. (2010), $\boldsymbol{\alpha}$ can be viewed as pseudo-counts of observations for each class before actual data collection. This interpretation provides a natural mechanism for incorporating existing knowledge about age distributions, such as from census data or previous mortality studies (Li et al., 2020; Mulick et al., 2021). The strength of the prior information is determined by the magnitude of $\boldsymbol{\alpha}$. Large values of $\boldsymbol{\alpha}$ correspond to strong prior knowledge, causing the Dirichlet-multinomial distribution to approximate the multinomial distribution closely. Conversely, small values represent weak or minimal prior information. This flexibility allows researchers to appropriately weight the influence of prior knowledge depending on its reliability and relevance to the current study population.

Recent methodological advances in Bayesian analysis of demographic data have demonstrated the importance of prior specification when dealing with sparse or partially ob-

served data (Schumacher et al., 2020; Wu et al., 2021a). Our theoretical results provide formal justification for these practices and establish specific conditions under which incorporating partially classified data improves estimation accuracy. In the context of global health monitoring, where data quality and availability vary substantially across regions, our approach offers a principled framework for maximizing the utility of available data while accounting for its limitations. The decision-theoretic justification provides a formal basis for integrating data across different reporting systems and classification schemes, potentially improving the accuracy and reliability of vital statistics in resource-limited settings.

2.3.4 Gibbs Sampling Scheme with Overlapping Partial Classifications

In practical applications of ACSU5M studies, data sources often employ inconsistent age categorization schemes that do not follow a strictly nested hierarchy (Diaz et al., 2021). This creates a more complex scenario where age groups A_j may overlap with each other, introducing dependencies in the posterior distribution that make direct analytical solutions intractable. Unlike the disjoint case discussed in Section 2.3.2, where closed-form Bayes estimators are available, the overlapping case requires computational approaches to approximate the posterior distribution. For example, one data source might report deaths in age groups 0-6 months and 6-59 months, while another uses the categories 0-12 months and 12-59 months, creating overlapping classifications at 6-12 months.

To address these challenges, we adopt a data augmentation approach (Tanner and Wong, 1987) by introducing latent count variables that decompose the partially classified observations into their constituent components. Specifically, let $((Z_{i|A_j})_{i=1}^{k})_{j=0}^{k}$ denote the unobserved counts of observations that belong to category i and are included when counting for group A_j. These latent variables satisfy the constraint:

$$Y_{A_j} = \sum_{i=1}^{k} Z_{i|A_j}.$$

This decomposition establishes a hierarchical structure where the observed counts Y_{A_j} are aggregations of the latent counts $Z_{i|A_j}$. The augmented likelihood becomes:

$$P(Y_{A_j}|\boldsymbol{\theta}) = \sum_{\{Z_{i|A_j}:\sum_i Z_{i|A_j}=Y_{A_j}\}} P(\{Z_{i|A_j}\}|\boldsymbol{\theta}).$$

While this augmented model remains complex, it enables us to implement an efficient MCMC estimation approach using Gibbs sampling (Gelfand, 2000; Geman and Geman, 1984). The Gibbs sampling algorithm alternates between sampling from the conditional distributions of the latent variables and the model parameters. For our overlapping partial classification problem, the algorithm proceeds as follows for the t-th iteration:

1. For each group $j = 0, 1, \ldots, m$:

$$(Z_{i|A_j})_{i=1}^k \sim \text{Multi}(y_{A_j}, \boldsymbol{\pi}_j),$$

   where $\boldsymbol{\pi}_j = (\theta_i \mathbb{1}\{i \in A_j\}/\theta_{A_j})_{i=1}^k$ represents the conditional probabilities of each category i within group A_j.

2. Compute the augmented counts:

$$c_i^{(t)} = x_i + \sum_{j=0}^{m} z_{i|A_j}.$$

3. Sample the updated parameter vector:

$$\boldsymbol{\theta}^{(t+1)} \sim \text{Dir}(\boldsymbol{\alpha}_t),$$

   where $\boldsymbol{\alpha}_t = (\alpha_i + c_i^{(t)})_{i=1}^k$ combines the prior hyperparameters with the augmented counts.

This algorithm leverages the conjugacy between the Dirichlet prior and multinomial likelihood to facilitate efficient sampling (Bernardo and Smith, 2009). The conditional

distributions in the Gibbs sampler are derived from the joint posterior of the augmented model:

$$P(\boldsymbol{\theta}, \{Z_{i|A_j}\}|\mathbf{X}, \mathbf{Y}) \propto P(\boldsymbol{\theta}|\boldsymbol{\alpha})P(\mathbf{X}|\boldsymbol{\theta})P(\{Z_{i|A_j}\}|\boldsymbol{\theta}).$$

The Gibbs sampling approach for overlapping classifications inherits the theoretical guarantees of MCMC methods. Under standard regularity conditions, the Markov chain constructed by the algorithm converges to the true posterior distribution (Roberts and Rosenthal, 2004). The rate of convergence depends on the complexity of the overlap structure and the dimension of the parameter space. Recent developments in computational statistics have enhanced the efficiency of Gibbs sampling for complex hierarchical models. Techniques such as parameter expansion (Liu and Wu, 1999), adaptive MCMC (Roberts and Rosenthal, 2009), and Hamiltonian Monte Carlo (Hoffman et al., 2014) can be incorporated to improve mixing and convergence properties. These advancements are particularly relevant for large-scale ACSU5M studies with numerous overlapping age categories. Hierarchical Bayesian approaches with data augmentation provide a flexible framework for addressing complex data integration challenges in demographic research. Our Gibbs sampling scheme aligns with this perspective, offering a principled solution to the overlapping classification problem.

Practical implementation of the Gibbs sampling scheme requires attention to several computational aspects. The algorithm's efficiency depends on the initialization strategy, convergence diagnostics, and the number of iterations (Gelman and Shalizi, 2013). Modern Bayesian software packages such as JAGS (Hornik et al., 2003), Stan (Carpenter et al., 2017), and PyMC (Salvatier et al., 2016) offer robust frameworks for implementing the Gibbs sampling algorithm described above.

2.4 Numerical Studies

In this section, we present a comprehensive analysis of the finite-sample performance of the proposed estimators through extensive numerical experiments. This investigation

aims to empirically validate our theoretical findings regarding the dominance relationships between the estimators $\hat{\boldsymbol{\theta}}$ and $\tilde{\boldsymbol{\theta}}$ under various data scenarios and classification settings. Our simulation studies are designed to explore the effects of sample sizes, partition structures, and prior distributions on estimation accuracy.

2.4.1 Simulation Settings

We considered three distinct classification scenarios with varying levels of complexity:

- *Scenario (i)*: $S = \{1, 2, 3\}$, $A_0 = \{1\}$, $A_1 = \{2, 3\}$.
- *Scenario (ii)*: $S = \{1, 2, \ldots, 9\}$, $A_0 = \{1, 2, 3, 4\}$, $A_1 = \{5, 6, 7, 8, 9\}$.
- *Scenario (iii)*: $S = \{1, 2, 3\}$, $A_0 = \{1, 2\}$, $A_1 = \{2, 3\}$.

The first two scenarios represent cases where the partial classifications form disjoint partitions of the complete categorization, with *Scenario (i)* featuring a simple three-class problem and *Scenario (ii)* involving a more complex nine-class classification task. In contrast, *Scenario (iii)* introduces overlapping partial classifications, where category 2 appears in both A_0 and A_1, representing a common challenge in real-world data integration problems. For each scenario, we employed Jeffrey's prior (Jeffreys, 1946) for the multinomial parameters, setting $\boldsymbol{\theta}_1 = (1/3, 1/3, 1/3) \in \mathbb{R}^3$ for *Scenarios (i)* and *(iii)*, and $\boldsymbol{\theta}_2 = (1/9, \ldots, 1/9) \in \mathbb{R}^9$ for *Scenario (ii)* .

To thoroughly investigate the impact of varying sample sizes on estimation performance, we designed a two-dimensional grid of experimental conditions:

- *Condition (i)*: Fixed completely classified sample size $N \in \{50, 100, 200\}$ with varying partially classified sample size $N' \in \{50, 100, 200, 500, 1000, 2000\}$.
- *Condition (ii)*: Fixed partially classified sample size $N' \in \{50, 100, 200\}$ with varying completely classified sample size $N \in \{50, 100, 200, 500, 1000, 2000\}$.

These experimental conditions enable us to examine both the effect of increasing the amount of partially classified data while keeping the completely classified data constant (*Condition (i)*) and the impact of incorporating more completely classified data while maintaining a fixed amount of partially classified information (*Condition (ii)*). This comprehensive design provides insights into the relative importance of each data type for estimation accuracy.

2.4.2 Performance Evaluation with Disjoint Partitions

Figures 2.1 and 2.2 display the Bayes risk of $\tilde{\boldsymbol{\theta}}$ (estimated using only the completely classified data) and $\hat{\boldsymbol{\theta}}$ (estimated with additional partially classified data) for *Scenario (i)* and *Scenario (ii)*, respectively. Defined as the expected loss under the posterior distribution, the Bayes risk provides a comprehensive measure of estimation accuracy that accounts for both bias and variance components (Berger and Berger, 1985).

Several key observations emerge from our analysis of these results:

1. *Effect of classification granularity:* As we increase the disaggregation level k while holding other parameters constant, the risk of both $\hat{\boldsymbol{\theta}}$ and $\tilde{\boldsymbol{\theta}}$ increases. This pattern aligns with statistical theory, as finer-grained classification requires estimating more parameters with the same amount of data, leading to increased estimation uncertainty. Specifically, the risk in *Scenario (ii)* ($k = 9$) is generally higher than in *Scenario (i)* ($k = 3$).

2. *Influence of theoretical conditions:* In *Scenario (i)* , where the sufficient condition for the dominance of $\hat{\boldsymbol{\theta}}$ is not satisfied ($\min_j \alpha_{A_j} < 2$), we observe instances where $\tilde{\boldsymbol{\theta}}$ achieves lower risk than $\hat{\boldsymbol{\theta}}$ when N' is relatively small. This finding empirically validates our theoretical result in Theorem 1, highlighting that the incorporation of partially classified data does not always lead to improved estimation when the sufficient condition is violated.

3. *Consistent improvement under sufficient conditions:* In contrast, for *Scenario (ii)* where the sufficient condition is satisfied ($\min_j \alpha_{A_j} \geq 2$), $\hat{\boldsymbol{\theta}}$ consistently outperforms $\tilde{\boldsymbol{\theta}}$ across all sample size configurations, even when N' is small. This demonstrates the practical utility of our theoretical conditions as guidelines for when to incorporate partially classified data.

4. *Sample size effects:* When fixing N' and increasing N (second row of each figure), the risk of both estimators decreases monotonically. This decrease is particularly dramatic as N grows large, reflecting the diminishing marginal value of partially classified data when abundant completely classified data is available. This trend aligns with asymptotic theory, which suggests that estimators converge to the actual parameters as the sample size increases.

The relative improvement of $\hat{\boldsymbol{\theta}}$ over $\tilde{\boldsymbol{\theta}}$ varies depending on the specific configuration of sample sizes and partition structures. Notably, the advantage of incorporating partially classified data tends to be more pronounced when N is small relative to N', suggesting that partially classified data offers the most significant benefit in scenarios where completely classified data is limited or expensive to obtain.

2.4.3 Performance Evaluation with Overlapping Partitions

When partial categorizations are not disjoint, as in *Scenario (iii)*, the estimation problem becomes more complex. Figure 2.3 presents the Bayes risk of $\hat{\boldsymbol{\theta}}_1$ and $\tilde{\boldsymbol{\theta}}_1$ for this scenario, estimated with Gibbs sampling. Our analysis of the overlapping partition scenario reveals the following findings:

1. *Lack of guaranteed improvement:* Unlike in the disjoint partition scenarios, there are no theoretical guarantees that incorporating partially classified data will improve estimation accuracy when partitions overlap. Indeed, our empirical results confirm

this theoretical gap, as we observe that $\hat{\boldsymbol{\theta}}$ frequently exhibits higher risk than $\tilde{\boldsymbol{\theta}}$, particularly when N' is large relative to N.

2. *Risk dynamics with increasing N':* With fixed N, the risk of $\hat{\boldsymbol{\theta}}$ initially decreases but then increases as N' grows very large. This non-monotonic behavior suggests that there may be an optimal amount of partially classified data to incorporate, beyond which additional partially classified data introduces more noise than signal into the estimation process.

3. *Risk dynamics with increasing N:* When N' is fixed and N increases, the risk of $\hat{\boldsymbol{\theta}}$ consistently decreases. This pattern indicates that incorporating more completely classified data helps mitigate the uncertainties introduced by the overlapping structure in the partially classified data, aligning with the intuition that high-quality data can compensate for the complexities introduced by lower-quality data.

The deterioration in performance with large N' in the overlapping case can be attributed to the increased model complexity and potential for conflicting information between the overlapping categories. When category 2 appears in both A_0 and A_1, the model must reconcile potentially inconsistent signals about this category from different partial classifications, leading to increased estimation uncertainty.

2.5 Empirical Applications

To assess the practical utility of our data disaggregation approach in real-world settings, we conducted two empirical case studies using data from well-established public health surveillance systems:

1. *Sample Registration System (SRS):* The SRS provides continuous enumeration of births and deaths in a nationally representative sample of population clusters. These systems are particularly valuable in countries where vital registration systems are incomplete (Sethi et al., 2008).

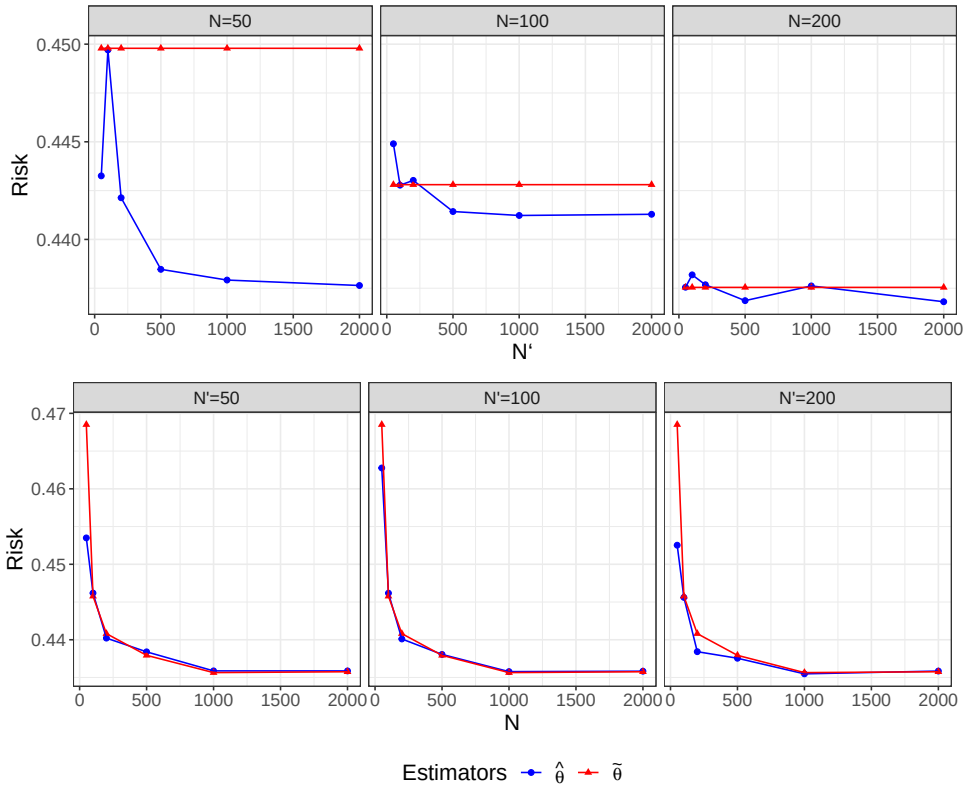


Figure 2.1: Risk comparison between $\hat{\boldsymbol{\theta}}_1$ and $\tilde{\boldsymbol{\theta}}_1$ estimators in *Scenario (i)*. Bayes risk is shown with N fixed and N' increasing (top row) and N' fixed and N increasing (bottom row).

2. *Demographic and Health Surveys (DHS):* The DHS program collects and disseminates accurate, nationally representative data on health and population in developing countries through household surveys (Corsi et al., 2012).

2.5.1 Maternal and Child Health Surveillance System Data of China

This section presents an empirical validation of our Bayesian age reconciliation methodology using the Maternal and Child Health Surveillance System (MCHSS) dataset of China. The MCHSS dataset spans two decades ($1996 - 2015$) and contains detailed mortality records for children under five years of age within designated surveillance areas. Death records in this dataset are systematically categorized into six distinct age strata:

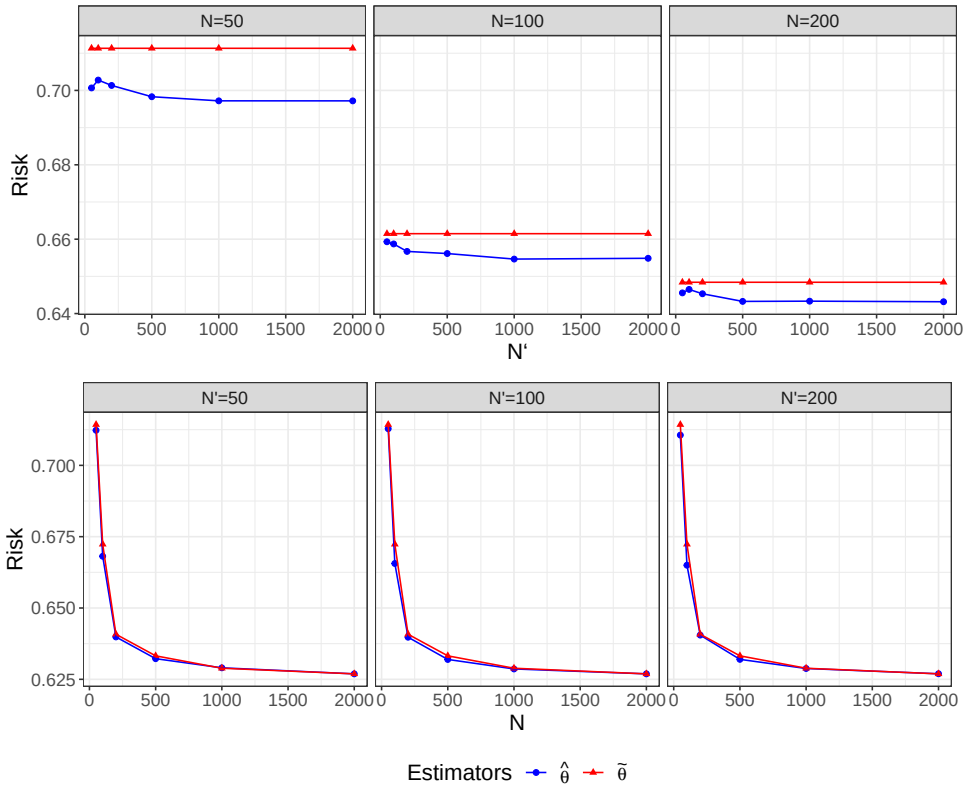


Figure 2.2: Risk comparison between $\hat{\boldsymbol{\theta}}_2$ and $\tilde{\boldsymbol{\theta}}_2$ estimators in *Scenario (ii)*. Bayes risk is shown with N fixed and N' increasing (top row) and N' fixed and N increasing (bottom row).

$0-6$ days, $7-27$ days, $1-5$ months, $6-11$ months, $12-23$ months, and $24-59$ months, which we index as $S = \{1, 2, 3, 4, 5, 6\}$. Each death is also classified into one of eight non-overlapping, exhaustive categories of CODs.

To rigorously evaluate our age disaggregation methodology under a nested age structure, we constructed a synthetic dataset derived from the MCHSS records collected between 1996 and 2005. In this synthetic dataset, we deliberately structured the data such that ages are partially classified into three broader groups: (i) $0-27$ days, (ii) $1-11$ months, and (iii) $12-59$ months. This partial classification simulates the typical scenario in demographic and health surveys where complete age disaggregation is unavailable.

For our empirical validation, we leveraged the MCHSS data, which includes six age

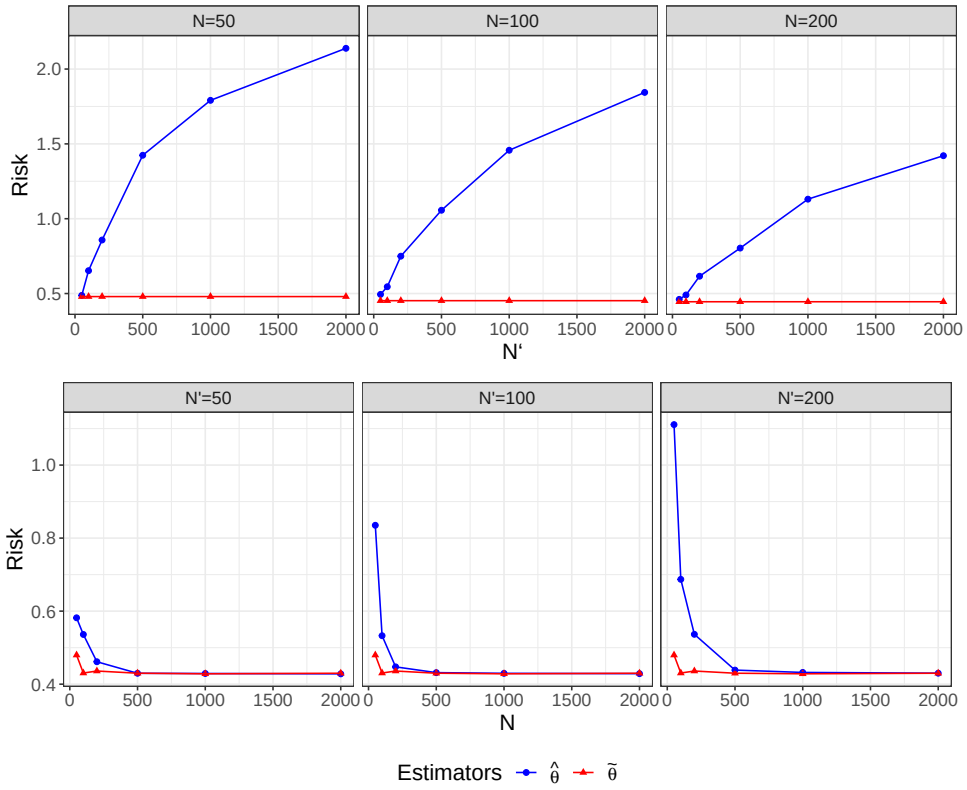


Figure 2.3: Risk comparison between $\hat{\boldsymbol{\theta}}_1$ and $\tilde{\boldsymbol{\theta}}_1$ estimators in *Scenario (iii)* . Bayes risk is shown with N fixed and N' increasing (top row) and N' fixed and N increasing (bottom row).

groups collected from 2006 to 2015, as the fully classified data. Our objective was to apply our Bayesian age reconciliation method to perform age group disaggregation on the partially classified data with three age groups, which we index as $A_1 = \{1, 2\}$, $A_2 = \{3, 4\}$, and $A_3 = \{5, 6\}$. Our goal is to evaluate the method's performance by comparing the disaggregated results with the actual fine-grained age-cause distribution from the original data from 1996 to 2005.

We evaluated the effectiveness of our proposed Bayesian age reconciliation method by comparing its predicted results to the actual MCHSS data from 1996 to 2005, as shown in Table 2.2. For illustrative purposes, the raw death counts estimated using the posterior predictive distribution are rounded to the nearest integers. Our quantitative assessment

True 1996-2005 MCHSS Data

	1	2	3	4	5	6
1	3550	638	138	9	2	0
2	4398	183	24	0	0	0
3	1783	568	732	373	264	270
4	95	78	459	224	268	473
5	732	231	463	176	584	1333
6	39	92	501	361	270	196
7	1227	991	1671	570	428	328
8	1051	722	491	269	230	261

Predicted 1996-2005 MCHSS Data

	1	2	3	4	5	6
1	3474	714	145	2	1	1
2	4258	323	23	1	0	0
3	1659	692	740	365	291	243
4	101	72	454	229	274	467
5	572	391	471	168	667	1250
6	26	105	536	326	321	145
7	1094	1124	1635	606	435	321
8	1059	714	515	245	243	248

Table 2.2: Comparison of true and disaggregated MCHSS data using Bayesian age reconciliation for $1996 - 2005$. Age groups are indexed by columns $(1 - 6)$ and COD by rows $(1 - 8)$.

reveals that the Bayesian age reconciliation method successfully recovers 93.0% of the death counts in their correct cross-classifications. To contextualize this performance, we compared our approach with a simple data integration method using random assignment, which achieved only 52.6% correct classification. This substantial improvement of more than 40% points demonstrates the robust capability of our proposed method to reconstruct age-disaggregated mortality patterns from partially observed data accurately.

The performance of our method is particularly noteworthy given the heterogeneous distribution of deaths across age groups and causes. For instance, causes predominantly

affecting neonates (COD 1 and 2) show a high concentration in the age group $0-6$ days, while other causes, such as injuries (COD 5), demonstrate different age patterns. Our method successfully captures these complex age-cause interaction patterns.

Impact on Mortality Estimation

To comprehensively examine the impact of utilizing the disaggregated data on the estimation of age- and cause-specific child mortality, we performed age disaggregation cross-classified by region $r \in \{1, \ldots, R = 6\}$, year $t \in \{1, \ldots, T = 10\}$, and COD $c \in \{1, \ldots, C = 8\}$. We then applied the hierarchical Bayesian model framework proposed by Schumacher et al. (2020) to estimate child mortality rates from both the true and the disaggregated 1996-2005 MCHSS data.

The model structure is formulated as:

$$\begin{aligned} y_{r,a,t,c} | N_{r,a,t}, \lambda_{r,a,t,c} &\sim \text{Poisson}(N_{r,a,t}\lambda_{r,a,t,c}) \\ \log(\lambda_{r,a,t,c}) &= \alpha + \beta_r^R + \beta_a^A + \beta_c^C + \beta_{a,c}^{AC} + \beta_{r,c}^{RC} + \beta_{a,r}^{AR} \\ &\quad + \gamma_{r,a^*[a],c^*[c]}(t) + \epsilon_{r,a,t,c}. \end{aligned}$$

In this model, $N_{r,a,t}$ represents the person-years exposure while $y_{r,a,t,c}$ denotes the death counts attributable to cause c within region r and age group $a \in \{1, \ldots, A = 6\}$. Given that deaths are rounded to the nearest integer, the model employs a Poisson likelihood function. The overall intercept is represented by α, which establishes the baseline level across all observations. Fixed effects are incorporated through β_r^R, β_a^A, and β_c^C, which account for region-specific, age-specific, and cause-specific effects, respectively. First-order interaction terms capture the relationships between pairs of factors. Specifically, $\beta_{a,c}^{AC}$ represents the interaction between age and cause, $\beta_{r,c}^{RC}$ captures the region-cause interaction, and $\beta_{a,r}^{AR}$ accounts for the age-region interaction effects. The temporal component is modeled through $\gamma_{r,a^*[a],c^*[c]}(t)$, which serves as a random effect on time following a second-order random walk distribution. This specification enables the modeling of smooth time trends that vary simultaneously across region, age group,

and cause dimensions. Finally, $\epsilon_{r,a,t,c}$ represents the residual error term, capturing the remaining unexplained variation in the model after accounting for all main effects and interactions. The indices $a^*[a]$ and $c^*[c]$ group ages and causes into broader categories to enhance model stability, following the approach described in Schumacher et al. (2020). The model incorporates structured priors for the fixed and random effects, with hyperparameters estimated from the data using a fully Bayesian approach. This model specification allows us to estimate age-, cause-, and region-specific mortality rates while appropriately accounting for interactions and temporal trends. The comparison between models fitted to the true data versus the disaggregated data provides insights into the utility of our age reconciliation method for downstream epidemiological analyses.

Figures 2.4 present the posterior medians and credible intervals for the estimated log mortality rates in the early neonatal ($0-6$ days) and late neonatal ($7-27$ days) age groups over the period $1996-2005$ for non-communicable diseases in the eastern urban region of China. To disentangle the contributions of different model components to these estimates, we display results for two model specifications: a reduced model with fixed effects only, and the full model incorporating both fixed effects and random effect error terms, as proposed by Schumacher et al. (2020). Both the true and disaggregated data yield models that adequately capture the stochastic variation present in the empirical data. The consistency between the fitted models demonstrates the robustness of our age disaggregation methodology. However, we observe certain systematic differences that warrant attention. Specifically, the estimated log mortality rates in the $0-6$ days age group derived from the disaggregated data are consistently higher than those from the true data across the time series. This upward bias likely stems from misclassification during the disaggregation process, where a proportion of deaths may have been incorrectly assigned to the age group $0-6$ days rather than the age group $7-27$ days.

Despite these discrepancies, the model fitted to the disaggregated data successfully captures the overall age-specific temporal trends observed in the actual data. This success can be attributed to the hierarchical structure of the model, which enables information

sharing (borrowing strength) across different age strata, regions, and causes of death. This feature is particularly valuable when working with sparse data, as is often the case with ACSU5M data in many LMICs. The credible intervals derived from the disaggregated data are slightly wider than those from the actual data, reflecting the additional uncertainty introduced by the age reconciliation process. Nevertheless, the overlapping credible intervals between the two models suggest that the differences are not statistically significant at the conventional $\alpha = 0.2$ level, reinforcing the validity of our approach.

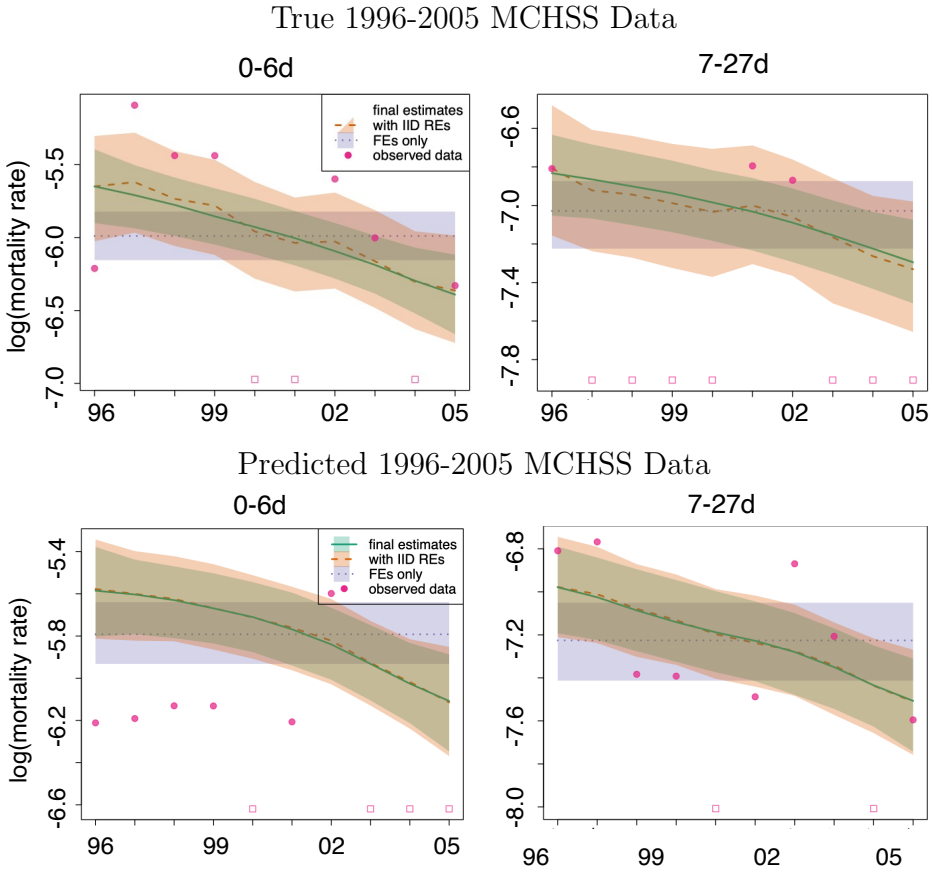


Figure 2.4: Log mortality rate estimates for non-communicable diseases in the east urban region using MCHSS data. Points show empirical data, lines indicate posterior medians, and shaded areas represent credible intervals. Open squares denote age-cause combinations with zero observed deaths.

Since the hierarchical Bayesian model of Schumacher et al. (2020) also estimates

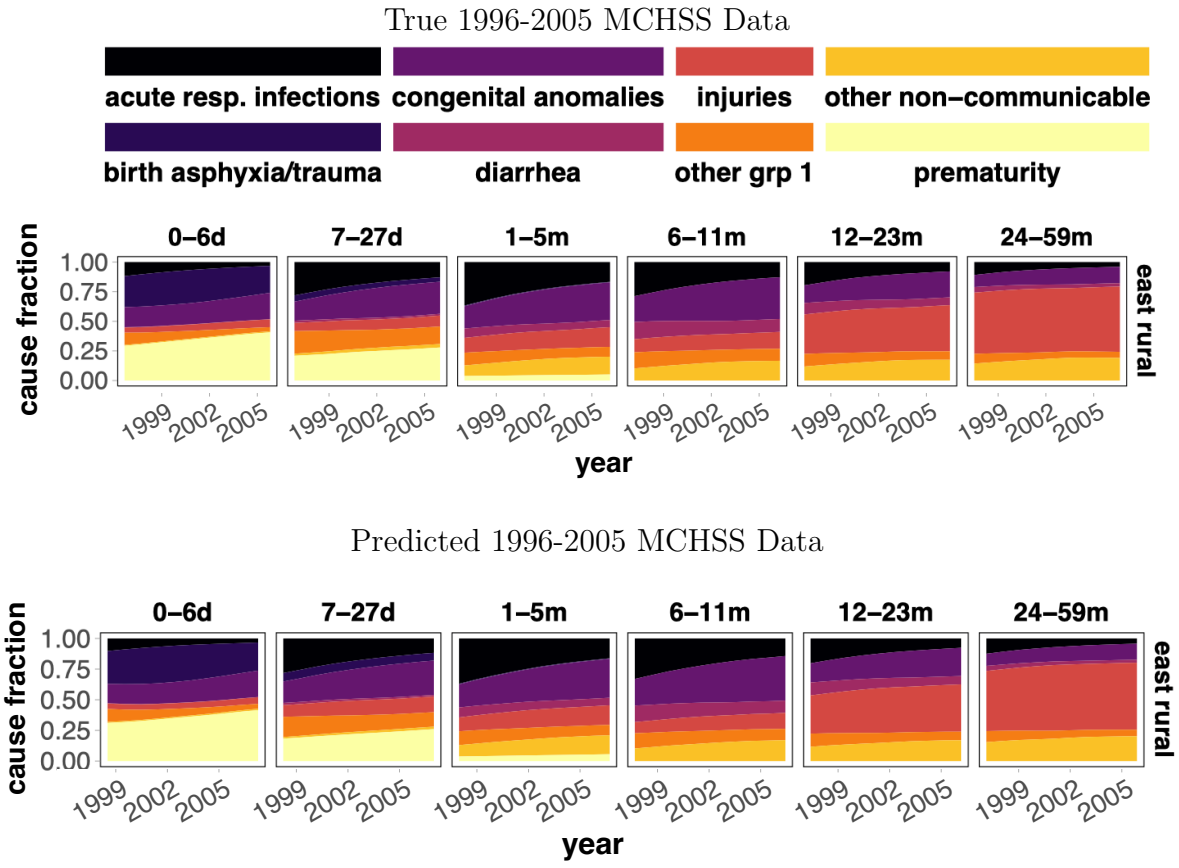


Figure 2.5: Comparison of estimated CSMFs between models using true and estimated MCHSS data for the east rural region. Both models show consistent temporal trends and CSMF estimates. The legend indicates the eight COD categories reported in MCHSS.

cause-specific child mortality fractions (CSMFs), we conducted a comparative analysis of the estimated CSMFs between models fitted to the true and disaggregated MCHSS data. Figure 2.5 illustrates this comparison for the eastern rural region of China across the study period. The temporal evolution of CSMFs estimated from the disaggregated data closely mirrors that of the true data, with remarkable concordance in both the relative proportions and time trends for major causes of death. Both models capture the gradual decline in infectious diseases and concomitant increase in the relative burden of non-communicable conditions over the decade, consistent with China's epidemiological

transition during this period (Song et al., 2016).

To further assess the robustness of our approach across different demographic and epidemiological contexts, we extended our comparative analysis to include all six regions covered by the MCHSS: *east urban*, *east rural*, *mid urban*, *mid rural*, *west urban*, and *west rural*. This comprehensive evaluation allows us to determine whether the performance of our age disaggregation method varies systematically by urbanicity or economic development level, which could have important implications for its application in diverse settings. Our extended analysis (documented in Appendix A) revealed generally consistent performance across all regions, with slight variations in accuracy. The method performed marginally better in urban regions compared to rural areas, potentially due to the higher completeness and quality of mortality registration in urban settings Wang et al. (2007); He et al. (2017). Despite these nuanced variations, the overall consistency of our method's performance across diverse contexts underscores its utility for harmonizing age-disaggregated mortality data at the national and subnational levels.

2.5.2 Bangladesh Demographic and Health Survey Data

The Bangladesh Demographic and Health Survey (BDHS) represents one of the most comprehensive population-based health surveys conducted in Bangladesh, providing nationally representative data on demographic and health indicators (NIPORT, 2012). For our study, we utilized verbal autopsy (VA) data collected through the BDHS in two distinct periods: 2011 and $2017 - 2018$ (NIPORT and ICF, 2016).

The VA component of the BDHS follows standardized protocols established by the World Health Organization (WHO) for collecting information on deaths occurring outside medical facilities (Thomas et al., 2018). For each reported death in the survey households, trained interviewers administered detailed VA questionnaires to close relatives or caregivers of the deceased. These questionnaires captured comprehensive information about symptoms, signs, and circumstances preceding the death (Kalter et al., 2011). Following data collection, rigorous physician reviews were conducted to determine the case-by-case

COD. This process involved independent assessment by at least two qualified physicians, with a third reviewer resolving any discrepancies (Streatfield et al., 2014). Through this methodology, 11 common CODs were identified and recorded across the surveys, providing a standardized classification system for mortality analysis. The age at death for each individual was also meticulously documented, allowing for age-specific mortality analysis.

Experimental Design for Non-nested Age Structure Evaluation

To rigorously evaluate the performance of our proposed Bayesian age reconciliation method under conditions of non-nested age structures, we designed a comprehensive experimental framework using the BDHS data. This experimental design aimed to simulate real-world scenarios where data sources report mortality information using overlapping but non-identical age categorizations.

First, we processed the $2017-2018$ BDHS data to create standardized age-disaggregated data with six distinct age groups: $0-6$ days, $7-27$ days, $1-5$ months, $6-11$ months, $12-23$ months, and $24-59$ months, indexed as $S = \{1, 2, 3, 4, 5, 6\}$. These fine-grained age classifications served as our fully classified reference data, representing the target level of age disaggregation that health monitoring systems typically aspire to achieve (Diaz et al., 2021). Subsequently, we constructed a synthetic dataset based on the 2011 BDHS data with three age categories deliberately structured to create overlapping groups: (i) $0-3$ months, (ii) $4-11$ months, and (iii) $12-59$ months, indexed as $A_1 = \{1, 2, 3\}$, $A_2 = \{3, 4\}$ and $A_3 = \{5, 6\}$. This configuration presents a particularly challenging scenario for age reconciliation methods, as the overlap between categories creates inherent ambiguity in the classification process. Specifically, deaths in the age group $1-5$ months (index 3) could potentially belong to either the $0-3$ months or $4-11$ months categories, necessitating a sophisticated probabilistic approach to resolve this uncertainty.

The non-nested structure we employed represents a common challenge in integrating demographic and health data, where different surveys, surveillance systems, or administrative records often employ disparate age classification schemes (Mikkelsen et al., 2015;

Diaz et al., 2021). Table 2.3 presents a comprehensive comparison between the actual and predicted 2011 BDHS data at the disaggregated level. The table displays the cross-tabulation of death counts by the six disaggregated age groups and eleven causes of death, allowing for a detailed assessment of classification accuracy at the cell level. Our analysis reveals that the proposed Bayesian age reconciliation method achieved an overall prediction accuracy of 70.4% at the cell level, representing a substantial improvement over the baseline random assignment approach, which yielded an accuracy of only 54.9%. This 15.5 percentage point improvement demonstrates the significant advantage of our model-based approach over naive allocation methods. Examining the results more granularly, we observe particularly strong performance in certain age-cause combinations. For instance, the method perfectly preserved the distribution of deaths for cause 3 (predominantly affecting older children) and cause 4 (concentrated in the neonatal period). These results suggest that our approach effectively captures and maintains strong age-cause association patterns. However, we also identify specific areas where the disaggregation performance was more modest. Notable misclassifications occurred primarily in cause 7, where some deaths were incorrectly reallocated from the $3-4$ age indices to index 3, and in cause 10, where the age distribution became more diffuse in the predicted data. These misclassifications predominantly occurred in cells where the original data exhibited complex age distributions that did not align clearly with either the $0-3$ months or $4-11$ months categories due to the overlapping nature of the age groups.

The performance variation across causes of death likely reflects differential age-specificity patterns. Causes with strong age predilections (e.g., neonatal conditions concentrated in the first week of life) were more accurately disaggregated than causes with more diffuse age distributions. This pattern aligns with theoretical expectations, as the model leverages the strength of age-cause associations to inform the disaggregation process. The empirical validation using both MCHSS and BDHS data demonstrates that our Bayesian age reconciliation method successfully preserves the essential mortality patterns in terms of both age-specific rates and cause distributions, while effectively addressing the chal-

lenge of inconsistent age groupings across data sources. It is worth noting that the BDHS data presents a more challenging test case than the MCHSS data examined in the previous section, due to both the non-nested age structure and the smaller sample size available for model training. The moderately lower accuracy achieved with the BDHS data (70.4% compared to 93.0% with MCHSS) reflects these additional challenges, yet still demonstrates the utility of our method even under suboptimal conditions.

2.6 Conclusions and Discussions

This chapter extends existing methods for estimating multinomial probabilities by developing a framework that accommodates data sources with different age group classifications. The primary innovation involves incorporating partially classified data, where ages are reported at varying levels of detail, into a unified Bayesian estimation approach. Unlike traditional methods that discard incomplete observations, this methodology utilizes information from all available sources to improve estimation accuracy. The work provides three theoretical contributions. First, it establishes the mathematical foundation for including partially classified data in multinomial estimation using the KL loss function. Second, it identifies conditions under which adding partially classified data improves Bayesian estimates. Third, it develops a computational framework that handles both hierarchical and overlapping age structures, making the approach applicable to various demographic and health research scenarios. Comprehensive testing through numerical simulations and real-world applications demonstrates considerable improvements over conventional methods in addressing age inconsistencies in ACSU5M studies. The method achieved classification accuracies of 93.0% with MCHSS data and 70.4% with BDHS data, compared to baseline accuracies of 52.6% and 54.9%, respectively. The performance difference between these applications reveals important method characteristics. The higher accuracy with the MCHSS data results from two factors: the hierarchical structure of age groups creates a more straightforward reconciliation problem, and the larger sample size provides more reliable information for parameter estimation. The

overlapping age categories in the BDHS data create inherent uncertainties, making reconciliation more challenging.

While our proposed age reconciliation method offers advantages in addressing age inconsistencies in under-five mortality studies, several promising directions necessitate further investigation. An immediate extension involves developing an integrated framework that simultaneously addresses misclassifications in both COD and age groups. Statistical algorithms commonly assign COD in demographic studies, particularly when using VA data from regions lacking comprehensive VR systems, though substantial uncertainties remain in these classification models (Mulick et al., 2021). A joint modeling approach using misclassification matrices within hierarchical Bayesian models would recognize the interdependence between age and cause classification errors, potentially yielding more accurate estimates of ACSU5M patterns. This approach would require careful specification of prior distributions that capture complex dependence structures between age and cause misclassifications, leveraging recent advances in structured prior specification and shrinkage priors. The computational challenges associated with estimating high-dimensional misclassification matrices would necessitate efficient MCMC algorithms or variational inference methods Moran et al. (2021); Blei et al. (2017). Another area for development involves establishing formal information-theoretic metrics that quantify the effects of data disaggregation. Our current methodology lacks these formal assessments, which would provide deeper insights into when and how partial classification information enhances estimation precision. Adapting Dempster-Shafer (DS) theory as an alternative framework for multinomial inference presents a promising direction, as this approach accommodates epistemic uncertainty arising from incomplete knowledge rather than random variation. The DS framework could quantify the informational contribution of partially classified data as belief functions, allowing researchers to evaluate the evidential weight provided by different data sources and their impact on final estimates (Zadeh, 1986; Sentz and Ferson, 2002b; Lawrence et al., 2009). Expanding the framework to address spatiotemporal dynamics represents another valuable improvement, particularly given the growing demand

for subnational mortality estimates that can inform targeted health interventions (Amek, 2013). The current framework primarily addresses age reconciliation at the national or regional levels. Still, extensions incorporating spatial correlation structures and temporal trends could improve the precision of disaggregated estimates in data-sparse areas by borrowing strength across adjacent geographical units and consecutive periods. Exploring multivariate spatial models that jointly model multiple causes of death could capture important dependencies in the spatial distribution of mortality causes, potentially improving both accuracy and coherence of subnational estimates (Yilema et al., 2022; Ahn et al., 2007). The integration of covariate information presents an additional opportunity for future research. Our current approach focuses primarily on reconciling age structures without explicitly incorporating demographic and socioeconomic factors such as maternal education, household wealth, and healthcare access that significantly influence both the level and age pattern of under-five mortality. Incorporating such covariates through regression-based extensions of our multinomial model, where probability parameters are linked to covariates through appropriate link functions, could enhance prediction accuracy, particularly in settings with limited data. Recent methodological developments in Bayesian additive regression trees (Chipman et al., 2010) and Gaussian process regression (Seeger, 2004; Rasmussen, 2003) offer flexible approaches for capturing complex, non-linear relationships between covariates and mortality patterns.

While the promising extensions outlined above remain to be explored, the current framework offers a practical and effective solution to ACSU5M estimation by developing a robust Bayesian framework for reconciling inconsistent age classifications across multiple data sources. Our approach, grounded in rigorous decision theory and validated through extensive empirical studies, demonstrates improvements in estimation accuracy compared to conventional approaches that discard incomplete data.

True 2011 BDHS data

	1	2	3	4	5	6
1	4	4	0	0	0	0
2	1	3	0	0	0	0
3	0	0	0	1	13	12
4	55	2	0	0	0	0
5	9	0	0	0	0	0
6	0	0	4	2	4	0
7	22	21	39	14	6	7
8	5	5	0	0	0	0
9	26	5	0	0	0	0
10	44	14	0	0	0	0
11	0	0	1	1	0	0

Predicted 2011 BDHS data

	1	2	3	4	5	6
1	4	3	1	0	0	0
2	2	1	1	0	0	0
3	0	0	0	1	13	12
4	55	2	0	0	0	0
5	8	1	0	0	0	0
6	0	0	4	2	2	2
7	23	13	57	3	2	11
8	7	3	0	0	0	0
9	26	4	1	0	0	0
10	31	23	4	0	0	0
11	0	1	1	0	0	0

Table 2.3: Evaluation of the proposed method for non-standard age categories using 2011 BDHS data. Comparison of actual versus predicted results shows strong overall performance with limited misclassification. Age groups are indexed by columns $(1-6)$ and causes of death by rows $(1-11)$.

Chapter 3

BAYESIAN ACTIVE QUESTIONNAIRE DESIGN FOR VERBAL AUTOPSIES

3.1 Introduction

Comprehensive mortality data, including information on cause-of-death (COD), provide essential insights into population health patterns and disease burdens across diverse contexts. Many low- and middle-income countries (LMICs) lack comprehensive vital registration (VR) systems, creating significant gaps in global mortality surveillance (AbouZahr et al., 2015). Consequently, approximately two-thirds of global deaths occur without registration or cause assignment (World Health Organization, 2021a), which limits evidence-based health policy development and resource allocation. Verbal autopsy (VA) has emerged as an essential tool to address this information gap. VA collects information through structured interviews with family members or caregivers who witnessed the circumstances surrounding a death, capturing essential diagnostic information when medical certification is unavailable (Thomas et al., 2018). The process involves trained interviewers administering standardized questionnaires that document symptoms, signs, risk factors, and circumstances that occurred before death (Soleman et al., 2006).

VA has been widely used across different health surveillance platforms, including demographic surveillance systems, national mortality monitoring programs, and multi-country research initiatives (Chandramohan et al., 2021). The collected data undergoes analysis through two primary methods: traditional physician review, where medical experts interpret interview responses to determine likely causes, or increasingly, computational algorithms that provide greater scalability (James et al., 2011). Early algorithmic approaches mainly adopted variations of the Naive Bayes (NB) classifier, which assumes

that symptoms exhibit conditional independence given underlying causes (Byass et al., 2019; McCormick et al., 2016a). Recent methodological advances have introduced more sophisticated Bayesian hierarchical models that better handle the complex relationships among symptoms and CODs (Kunihama et al., 2020a; Li et al., 2020; Moran et al., 2021). Beyond structured questionnaire analysis, researchers have explored natural language processing (NLP) approaches to extract diagnostic information from text-based narratives collected during VA interviews. However, evidence suggests that narrative-based approaches have not consistently performed better than analyses based solely on structured questionnaire data (Blanco et al., 2020), highlighting the continued value of systematic symptom documentation. A comprehensive examination of contemporary COD assignment methodologies can be found in Li et al. (2022).

While existing literature has focused primarily on automating COD classification processes, the actual data collection procedures have received considerably less research attention. This research gap is significant because one of the primary barriers to expanding VA implementation is the administration of complex and lengthy questionnaires (Surek-Clark, 2020). The current WHO 2016 standardized VA instrument contains 480 questions. Although respondents typically answer only a subset based on their specific responses, standard VA interviews still require completing $100-200$ items (Nichols et al., 2018). These extended interviews create substantial burdens for both respondents and interviewers. Recently bereaved family members may experience increased emotional distress from prolonged questioning, while interviewers can develop compassion fatigue from repeated exposure to grief narratives (Loh et al., 2021; Nichols et al., 2022; Hinga et al., 2021). These combined factors can lead to survey fatigue and reduced interview acceptance, which may compromise both data quality and population coverage.

Two notable approaches have been undertaken to systematically reduce the length of VA questionnaires over the past two decades. The first attempt by Serina et al. (2015a) examined the marginal associations between reported symptoms and causes of death. This approach identified and eliminated symptoms with lower diagnostic value based on

their statistical importance rankings. However, this work had a significant limitation in its reliance on a single, simplified classification algorithm, which meant the results were heavily influenced by the algorithm's parametric assumptions. More recently, the WHO developed the 2022 standard VA instrument, as described in Chandramohan et al. (2021) and Nichols et al. (2022). This initiative employed a model-agnostic evaluation framework to investigate symptom response patterns and their importance. Through mixed-methods analyses, researchers identified approximately 100 questions that could be removed without substantially reducing diagnostic accuracy.

Both previous approaches to shortening VA interviews remain limited by the fundamental structure of traditional survey instruments: the static questionnaire must comprehensively address symptoms associated with all potential CODs, regardless of their relevance to any particular case (Garenne, 2014). Since the COD represents the primary objective and remains unknown during the interview process, traditional approaches require uniform administration of all questions across all respondents. Consequently, each individual interview inevitably includes a substantial proportion of collected data that provides minimal relevant information for determining the specific COD under investigation. This occurs because many questions examine symptoms that are unrelated to the actual cause of the particular death being assessed.

This chapter presents a statistical framework for conducting adaptive VA interviews through dynamic question selection based on previously collected responses. The approach terminates interviews once sufficient diagnostic information has been obtained, optimizing the process for individual death classification rather than applying uniform reduction methods across all respondents. Drawing from active learning (AL) and computerized adaptive testing (CAT) literature, this work represents the first application of adaptive design principles specifically developed for VA contexts. While the implementation focuses on VA questionnaires, the proposed approach has potential applications for any survey instrument designed to classify respondents into predefined categories.

This research delivers several important contributions. First, we develop an active

question selection strategy that dynamically optimizes COD classification for individual cases. Empirical validation demonstrates that for most deaths, a small subset of actively selected questions achieves classification accuracy comparable to that of using the complete question set. Second, our adaptive questionnaire design is broadly applicable to existing probabilistic COD assignment algorithms, functioning independently of the specific analytical model used to characterize symptom-cause relationships. This model-agnostic approach provides greater flexibility in practice, allowing implementers to choose the most appropriate analysis framework while integrating our adaptive questionnaire design into their data collection protocols. Third, we extend existing work in psychometrics to a more rigorous Bayesian strategy that fully accounts for uncertainties inherent in COD classification models. This extension produces uncertainty-aware stopping rules that are better suited for high-stakes tasks, such as VA, where classification errors may significantly impact epidemiological understanding and subsequent public health interventions. Finally, we introduce a novel penalized variant of our adaptive questionnaire strategy that addresses practical constraints and interviewer preferences for question ordering. This innovation addresses implementation challenges in field settings, where specific question sequences may enhance respondent comprehension or improve interviewer workflow.

3.2 Preliminaries

3.2.1 Active Learning Paradigm

Active learning (AL) represents an important branch of machine learning research focused on optimizing data selection for model training. Rather than using all available data, AL algorithms strategically identify and select the most informative instances for inclusion in the learning process (Settles, 2011). This approach proves particularly valuable when labeling data requires significant human effort or expense, as it maximizes the value derived from limited labeling resources.

Previous research presents diverse query strategies for AL implementations. The most

prominent approaches include uncertainty sampling, which prioritizes instances where the current model exhibits maximum uncertainty (Lewis and Gale, 1994a); query-by-committee, which uses disagreements among an ensemble of models to identify informative instances (Seung et al., 1992); and approaches that aim to reduce variance (Cohn et al., 1996) or expected generalization error (Roy and McCallum, 2001). Each strategy offers distinct advantages depending on the specific learning context and classification objectives. Particularly relevant to our work are AL approaches that query complete feature vectors from partially observed datasets (Melville et al., 2004; Li and Oliva, 2021). This variant addresses scenarios where observations contain missing values across different feature dimensions, which is analogous to our VA problem, where responses to some questions remain unobserved until explicitly queried. AL has been successfully applied across diverse domains, including NLP (Kaushal et al., 2019), computer vision (CV) (Dor et al., 2020), and numerous other fields that require efficient data utilization. Applications of AL specifically to survey response collection remain relatively limited despite clear potential benefits. The most closely related precedent to our approach is the active matrix factorization approach for surveys measuring voter opinion proposed in Zhang et al. (2020). They employ an active question selection strategy to optimize the estimation of respondents' latent profiles under a low-rank matrix factorization model. While conceptually similar to our approach for VA questionnaires, their work addresses a fundamentally different objective—preference profiling rather than classification — and employs distinct mathematical formulations appropriate to that context.

3.2.2 Computerized Adaptive Testing

Computerized adaptive testing (CAT) represents a testing approach designed to identify and administer the optimal set of questions for each examinee. Unlike traditional fixed-form tests, CAT dynamically adjusts item selection based on individual response patterns, thereby improving measurement efficiency and precision in estimating examinees' latent traits (Weiss and Kingsbury, 1984). The conceptual foundations of CAT

were initially developed within item response theory (IRT) by Lord (1971), who introduced a probabilistic framework for adaptive item selection. In standard IRT-based CAT implementations, items are selected from a calibrated item bank to maximize test information at the current estimated ability, optimizing measurement precision where it is most needed for each individual (Van der Linden and Glas, 2010). A primary application of CAT occurs in cognitive diagnosis models (CDMs), referred to as cognitive diagnostic computerized adaptive testing (CD-CAT) (Cheng, 2009; Huebner, 2010). Unlike traditional IRT models, CDMs employ a discrete attribute space, characterizing examinees according to their mastery or lack of mastery of specific cognitive skills. Consequently, standard CAT approaches for IRT are not directly applicable to CDMs. Several CD-CAT methods have been proposed with different item selection strategies. Two of the most widely adopted approaches are the Shannon entropy method, proposed by Tatsuoka (2002a) and Tatsuoka and Ferguson (2003), which prioritizes items that maximize the expected reduction in classification uncertainty, and various procedures using KL information (Xu et al., 2003; Cheng, 2009), which identify items that best discriminate between competing attribute patterns. Alternative information metrics including mutual information (Wang, 2013a) and large deviation (Liu et al., 2015) have also been proposed for CD-CAT.

Our approach for VA questionnaire optimization draws inspiration from these KL information-based approaches in CD-CAT. We adapt these principles to the specific requirements of COD classification, creating a framework that strategically selects questions to maximize discriminative power between competing cause hypotheses based on accumulated evidence from previous responses.

3.3 Method

3.3.1 Bayesian Active Questionnaire Design

We assume that the complete VA question bank comprises J distinct questions, from which our adaptive algorithm selects optimal questions during the interview process. The response to question j for death i is denoted as X_{ij}, with the full response profile for that case represented as $X_i = \{X_{i1}, \ldots, X_{iJ}\}$. The COD is represented by the categorical variable $Y_i \in \{1, \ldots, C\}$. While we focus primarily on binary responses ($X_{ij} \in \{0, 1\}$) consistent with standard VA questionnaires, our framework can be extended to accommodate general response forms without modifying the active design formulation. The methodological basis of our approach relies on a previously established probabilistic model fitted on a training dataset $\{X_i, Y_i\}$ with $i = 1, \ldots, n$, which provides estimates for the joint distribution $p(X, Y)$ of symptoms and causes. This modeling framework enables statistical inference about symptom-cause relationships, forming the basis for informed question selection. Following the modeling framework, we adapt KL information procedures from CD-CAT literature (Cheng, 2009) to the VA context. During interview progression, after administering t questions (denoted by set $\mathcal{S}_t$), we systematically evaluate each remaining question $j \notin \mathcal{S}_t$ to identify which would provide the most discriminative information by having maximum distributional difference between the currently estimated cause and alternative possibilities (Tatsuoka, 2002b).

For a potential alternative cause y and the j-th question, we define the distributional divergence as:

$$D_j(\hat{y}_i^{(t)} \parallel y) = \sum_x q_j(x \mid \hat{y}_i^{(t)}) \log \left(\frac{q_j(x \mid \hat{y}_i^{(t)})}{q_j(x \mid y)} \right),$$

where $q_j(x \mid y) = p(X_{ij} = x \mid Y_i = y)$ represents the conditional distribution of the j-th indicator given that the COD is y, and $\hat{y}_i^{(t)} = \text{argmax}_y \, p(Y_i = y \mid \{X_{ij} : j \in \mathcal{S}_t\})$ denotes the estimated cause based on information collected at step t.

Among the various approaches proposed for aggregating KL distances across multiple

classification alternatives in the CD-CAT literature, we adopt the posterior weighted KL (PWKL) algorithm (Cheng, 2009). This approach uses divergence measures for all potential causes, weighted by their current posterior probabilities, to produce a comprehensive metric. For each candidate question j, we compute and maximize:

$$\text{Score}_j = \sum_{y=1}^{C} D_j(\hat{y}_i^{(t)} \parallel y) p(Y_i = y \mid \{X_{ij} : j \in \mathcal{S}_t\}). \tag{3.1}$$

Maximizing this weighted score ensures that questions are selected to provide the maximum diagnostic value for distinguishing between plausible causes, given the current state of knowledge.

The existing literature on CD-CAT typically assumes that conditional distributions for score computation are either precisely known or can be accurately estimated from data (Chang et al., 2019). This assumption does not hold in VA contexts, where these distributions must be derived from COD assignment models trained on limited data $\mathcal{D}$. Various Bayesian methods have been proposed to infer COD from VA data (McCormick et al., 2016a; Kunihama et al., 2020b; Li et al., 2021; Wu et al., 2021a), with observations demonstrating substantial uncertainty in both classification results and parameter estimations within these models.

To account for the complete posterior uncertainty of the conditional probabilities used in the PWKL score calculation, we propose the following posterior predictive PWKL score:

$$\text{PScore}_j = \int \text{Score}_j(\phi) p(\phi \mid \mathcal{D}) d\phi \approx \frac{1}{B} \sum_{b=1}^{B} \text{Score}_j(\phi^{(b)}), \tag{3.2}$$

where ϕ denotes all parameters in the assignment model, $\phi^{(b)}$ represents the b-th sample drawn from the posterior distribution $p(\phi|\mathcal{D})$, B is the total number of posterior samples, and $\text{Score}_j(\phi^{(b)})$ denotes the PWKL score evaluated at $\phi^{(b)}$.

Throughout the remainder of this chapter, we distinguish between two active question selection approaches: the "design using point estimates", which maximizes $\text{Score}_j(\hat{\phi})$ where $\hat{\phi}$ denotes the posterior mean of ϕ, and the "design using posterior predictive

scores", which maximizes PScore_j as defined above. Our formulation accounts for complete model uncertainty, representing a further extension of the modified PWKL score introduced by Kaplan et al. (2015), which only integrated latent classification uncertainty.

3.3.2 COD Assignment Model

The active selection methodology presented in the previous section demonstrates flexibility regarding the specific COD assignment model used for data analysis, requiring only that the conditional probabilities referenced in Equation (3.1) are computationally tractable. This model-agnostic approach enables practitioners to select frameworks that are optimally suited to their analytical contexts.

For illustration purposes, we implement a simplified adaptation of the algorithm introduced by McCormick et al. (2016a) to analyze the training dataset $\mathcal{D}$. Our analytical model assumes the following data generation process:

$$Y_i \sim \text{Cat}(\boldsymbol{\pi}),$$
$$p(X_i = x_i \mid Y_i = y) = \prod_j \theta_{yj}^{x_{ij}} (1 - \theta_{yj})^{1 - x_{ij}}.$$

We adopt conjugate priors $\theta_{yj} \sim \text{Be}(a_y, b_y)$ for the symptom probabilities conditional on causes and $\boldsymbol{\pi} \sim \text{Dir}(\boldsymbol{\alpha})$ for the cause prevalence vector. With this specification, the posterior distributions take the following analytical forms:

$$\boldsymbol{\pi} \mid \mathcal{D} \sim \text{Dir}(n_1 + \alpha_1, \ldots, n_C + \alpha_C),$$
$$\theta_{yj} \mid \mathcal{D} \sim \text{Be}\left(a_y + \sum_{i:Y_i = y} x_{ij},\ b_y + n_y - \sum_{i:Y_i = y} x_{ij}\right),$$

where $n_y = \sum_{i=1}^{n} \mathbf{1}(Y_i = y)$ represents the frequency of cause y across the sample space for $y = 1, \ldots, C$.

In scenarios involving partial COD information, where certain Y_i values are unknown in the data, the framework allows straightforward extension to generate posterior samples of $(\boldsymbol{\pi}, \boldsymbol{\theta}, Y_{miss})$, thereby accommodating missing data imputation.

3.3.3 Stopping Criterion

In adaptive testing frameworks where time constraints are critical, terminating the test after a predetermined number of questions represents a reasonable approach (e.g. Chen et al., 2012; Wang et al., 2011; Cheng, 2009). Such fixed-length termination rules can be readily incorporated into our adaptive VA questionnaire design. However, concluding the interview process only when sufficient precision is achieved generally constitutes a more appropriate practice.

Several approaches to early test termination rules are presented in the literature. Tatsuoka (2002a) introduced a stopping criterion based on the maximum probability of the classified category reaching a given threshold. This framework was later adapted by Hsu et al. (2013), who added a constraint on the second-highest probability value. In our notation, the Hsu et al. (2013) criterion suggests terminating the questionnaire when the maximum value of $p(Y_i = y \mid \{X_{ij} : j \in \mathcal{S}_t\})$ exceeds $p_{1\text{st}}$, while ensuring that the second-highest probability remains below $p_{2\text{nd}}$, where $p_{1\text{st}} \geq p_{2\text{nd}}$ represent two predetermined thresholds. While this stopping rule offers simplicity, reliance on point estimates of classification probabilities may lead to premature termination, particularly when parameter estimates exhibit substantial uncertainty. To address this limitation, we compute the posterior predictive probability of satisfying a predefined stopping criterion analogous to that proposed by Hsu et al. (2013). At each iteration, we evaluate $p_y^{(b)} = p(Y_i = y \mid \{X_{ij} : j \in \mathcal{S}_t\}, \phi^{(b)})$ for the b-th posterior sample $\phi^{(b)}$. We then identify the current most probable cause assignment $y^\star$ as the mode of $\{\text{argmax}_{y=1,\ldots,C}\, p_y^{(b)}\}_{b=1,\ldots,B}$.

Subsequently, we derive the posterior predictive probability for the event:

$$p(Y_i = y^\star \mid \{X_{ij} : j \in \mathcal{S}_t\}) > p_{1\text{st}},$$
$$p(Y_i = y \mid \{X_{ij} : j \in \mathcal{S}_t\}) < p_{2\text{nd}}, \quad \forall y \neq y^\star.$$

Questionnaire administration terminates when this computed probability exceeds a specified tolerance threshold $r \in (0, 1)$. Although we implement this particular stopping criterion in our analysis, the fully Bayesian nature of our proposed score formulation

readily accommodates other stopping criteria with minimal structural modification.

3.3.4 Question Flow in Active VA Questionnaire

Traditional VA instruments utilize carefully sequenced question structures that facilitate a natural conversational progression during interviews. In contrast, our proposed active questionnaire approach, while maximizing information gain efficiency, inherently disrupts this established flow by potentially transitioning between disparate topics related to mortality. Although this adaptive approach accelerates information acquisition, it introduces concerns regarding response validity when respondents navigate abrupt thematic transitions during emotionally sensitive interviews (Loh et al., 2021; Nichols et al., 2022). Recent ethnographic research on VA implementation has documented how the structure of questionnaires affects both respondent comfort and data quality, with poorly sequenced questions potentially triggering increased psychological distress and compromising response accuracy (Hinga et al., 2021).

To mitigate these concerns while preserving the efficiency benefits of active questioning, we implement deterministic branching logic by dynamically constraining the question search space $\mathcal{S}_t$ based on prior responses. This approach maintains critical structural dependencies within the VA instrument. Many standardized VA protocols contain conditional questions that become relevant only after specific prerequisite questions have been answered affirmatively. Our implementation addresses this by initializing $\mathcal{S}_0$ to contain only root-level questions, subsequently expanding $\mathcal{S}_t$ to include subordinate questions exclusively after their corresponding parent questions have received appropriate responses, thus preserving logical dependencies while optimizing information gain.

Beyond enforcing rigid conditional hierarchies, our framework accommodates softer structural preferences through a penalty-based approach that better reflects the expertise of interviewers and the needs of respondents. We extend the active design strategy by incorporating a question-transition penalty into our scoring mechanism. Formally, we

define a penalized posterior predictive score:

$$\text{PScore}'_j = \text{PScore}_j - \lambda D(j, j^{(t)}),$$

where $j^{(t)}$ represents the question index from iteration t, $\lambda > 0$ serves as a coefficient regulating transition penalization strength, and $D(j, k)$ constitutes a customizable distance function between questions j and k.

This formulation provides considerable implementation flexibility. For example, to encourage thematically cohesive questioning while allowing arbitrary sequencing within topical clusters, we can define $D(j, k) = 0$ for questions within the same thematic domain and $D(j, k) = 1$ otherwise. Through empirical tuning of λ, researchers can calibrate the balance between information optimization and interview coherence based on field-specific requirements and cultural contexts (Chandramohan et al., 2021).

3.4 Experiments

3.4.1 Synthetic Data Analysis

To evaluate our proposed active questionnaire method, we conduct a comprehensive analysis of synthetic data using two distinct modeling scenarios. These scenarios test performance under varying degrees of model specification accuracy. We examine two complementary data-generating mechanisms. The first scenario represents a well-specified case where we generate synthetic observations that align perfectly with the analytical framework described in Section 3.3.2. This ensures complete consistency with the underlying modeling assumptions. The simulation parameters include $C = 10$ CODs, $J = 50$ questions, uniform Dirichlet prior with $\boldsymbol{\alpha} = (1, \ldots, 1)$, and Beta hyper-parameters (a_y, b_y) set to $(0.5, 0.5)$ for causes $y \in \{1, 2, 3\}$, $(3, 3)$ for $y \in \{4, 5, 6\}$, and $(1, 3)$ for $y \in \{7, 8, 9, 10\}$. This parameterization creates realistic variation in symptom-cause associations. The second scenario introduces model misspecification through a more complex latent class structure. Under this framework, each COD consists of multiple unobserved subcategories. We introduce latent indicators $Z_i \mid Y_i = y \sim \text{Cat}(\boldsymbol{\lambda}_y)$ and define the response

mechanism as:

$$p(\mathbf{X}_i = \mathbf{x}_i \mid Y_i = y, Z_i = k) = \prod_j \theta_{ykj}^{x_{ij}} (1 - \theta_{ykj})^{1-x_{ij}},$$

where $\boldsymbol{\lambda}_y = (\lambda_{y1}, \lambda_{y2}, \lambda_{y3}) \sim \text{Dir}(1, 1, 1)$ and $\theta_{ykj} \sim \text{Beta}(1, 1)$ for all combinations of y, k, and j. This formulation introduces systematic model misspecification that tests the robustness of our AL approach under challenging conditions.

For each scenario, we generate training datasets of varying sizes with $n_{\text{train}} \in \{200, 1000\}$ to investigate the effects of sample size on performance. We maintain a fixed test set of $n_{\text{test}} = 200$ observations for consistent performance evaluation across all experimental conditions.

Fixed-Length Questionnaire Performance

Our initial evaluation examines classification performance under fixed questionnaire lengths, providing insights into the efficiency gains achievable through active question selection. Figure 3.1 presents the probability of correct COD classification across varying questionnaire lengths. The results demonstrate that both active selection strategies consistently achieve superior classification accuracy more rapidly than conventional sequential questioning approaches. Under the well-specified model condition, we implement an oracle baseline that utilizes true data-generating parameters. This comparison reveals that our active strategies approach oracle-level performance when sufficient training data is available.

Adaptive Stopping Rule Performance

We then investigate the performance implications of applying various adaptive stopping criteria to our active questionnaire design under identical data-generating conditions. We systematically compare classification accuracy across different stopping rules for both data-generating scenarios. Our evaluation considers three stopping rules. For

point estimate-based question selection, we implement the criterion established by Hsu et al. (2013). For posterior predictive scoring approaches, we examine stopping thresholds when $r = 50\%$ and 70% of posterior samples satisfy the termination criterion. The experimental design employs primary probability thresholds $p_{1\text{st}} \in \{0.8, 0.9\}$, with secondary thresholds computed as $p_{2\text{nd}} = (1 - p_{1\text{st}})/C + d(C - 2)(1 - p_{1\text{st}})/C$ for discrimination parameters $d \in [0, 1]$, following Hsu et al. (2013). Tables 3.1 and 3.2 present the comprehensive results. As expected, classification accuracy shows positive correlation with stricter primary thresholds $p_{1\text{st}}$ and more restrictive secondary thresholds $p_{2\text{nd}}$ (equivalently, smaller d values), reflecting the increased questioning required to satisfy more stringent termination criteria.

For illustrative purposes, we limit our presentation to results for $d \in \{0, 0.5\}$, although we also report questionnaire length distributions using median values and 5th/95th percentiles. A particularly notable finding emerges in Table 3.2: under model misspecification conditions, the $r = 70\%$ posterior probability stopping rule produces median questionnaire lengths comparable to those of point estimate approaches, while exhibiting extended upper tails in question counts. This conservative behavior results in enhanced overall accuracy, suggesting particular value in high-stakes applications. In clinical contexts such as VA, where misclassification carries significant consequences, conservative strategies that gather additional information for ambiguous cases may be preferable, particularly when the underlying analytical model exhibits specification uncertainty. Such scenarios favor stopping rules with larger r values in practical implementations.

Question Flow Performance

Finally, we examine scenarios where response accuracy deteriorates when questionnaire flow deviates from natural conversational sequences, reflecting realistic interviewing conditions. Denoting $j^{(t)}$ as the t-th question index, we simulate noisy responses through a

mechanism where:

$$X^*_{ij^{(t)}} = \begin{cases} 1 - X_{ij^{(t)}} & \text{with probability } \frac{D(j^{(t-1)}, j^{(t)})}{h}, \\ X_{ij^{(t)}} & \text{with probability } 1 - \frac{D(j^{(t-1)}, j^{(t)})}{h}. \end{cases}$$

We implement a normalized distance metric $D(j,k) = |j-k|/J$, ensuring that sequential questioning minimizes induced response noise. Our comparative analysis evaluates penalized scoring approaches (detailed in Section 3.3.4) using penalty parameters $\lambda \in \{2, 10\}$ under both low-noise ($h = 10$) and high-noise ($h = 2$) conditions, maintaining the data generating process from our initial experiments.

Figure 3.2 demonstrates that unpenalized active designs experience performance degradation in high-noise environments due to accumulated errors from non-sequential question ordering. In contrast, our penalized scoring methodology successfully mitigates these adverse effects, achieving classification accuracy that matches or exceeds optimally ordered static questionnaire designs. This finding confirms the practical importance of incorporating structural preferences into active learning frameworks for sensitive interview contexts.

3.4.2 Population Health Metrics Research Consortium Data Analysis

To evaluate the practical performance of our proposed approach, we conduct empirical validation using the Population Health Metrics Research Consortium (PHMRC) gold-standard VA dataset (Murray et al., 2011a). This dataset serves as a fundamental resource in the VA research community and represents the primary benchmark for validating computational COD assignment algorithms across diverse methodological approaches (McCormick et al., 2016b; Kunihama et al., 2020b; Moran et al., 2021).

The PHMRC data comprises 7841 adult mortality cases collected across six geographically and culturally diverse study sites: *Andhra Pradesh* and *Uttar Pradesh* in India; *Bohol Island* in the Philippines; *Dar es Salaam* and *Pemba Island* in Tanzania; and *Mexico City* in Mexico. The dataset's key strength lies in its gold-standard COD

p_{1st}	d	Stopping Rule	Acc	Median	Lower	Upper
0.8	0.5	Point Est	0.85	5	3	14
		Pred $r = 0.5$	0.92	10	4	50
		Pred $r = 0.7$	0.95	14	6	50
0.8	0	Point Est	0.96	8	5	35
		Pred $r = 0.5$	0.96	13	6	50
		Pred $r = 0.7$	0.96	20	7	50
0.9	0.5	Point Est	0.95	7	5	18
		Pred $r = 0.5$	0.95	12	6	50
		Pred $r = 0.7$	0.96	18	7	50
0.9	0	Point Est	0.97	10	7	50
		Pred $r = 0.5$	0.96	16	8	50
		Pred $r = 0.7$	0.96	23	8	50

Table 3.1: Classification accuracy and questionnaire length for synthetic data under the correctly specified model. Results show the median questionnaire length, along with the 5-th and 95-th percentiles, across different stopping rule conditions.

determinations, which were established through comprehensive clinical investigations including laboratory diagnostics, histopathological examinations, and advanced medical imaging modalities.

Our analysis framework operates on 34 distinct COD categories ($C = 34$), with questionnaire responses transformed into 168 binary symptom indicators ($J = 168$) following the standardized preprocessing pipeline established by Li et al. (2022). We examine the performance of using both fixed-length questionnaires and adaptive stopping criteria to provide a comprehensive performance assessment across different implementation scenarios. We employ 10-fold cross-validation to account for potential variability across different

p_{1st}	d	Stopping Rule	Acc	Median	Lower	Upper
0.8	0.5	Point Est	0.88	5	3	11
		Pred $r = 0.5$	0.88	4	3	18
		Pred $r = 0.7$	0.94	5	3	27
0.8	0	Point Est	0.96	6	5	17
		Pred $r = 0.5$	0.94	6	4	30
		Pred $r = 0.7$	0.99	8	5	50
0.9	0.5	Point Est	0.94	6	4	13
		Pred $r = 0.5$	0.96	6	4	23
		Pred $r = 0.7$	0.98	7	4	50
0.9	0	Point Est	0.96	7	6	20
		Pred $r = 0.5$	0.98	8	5	42
		Pred $r = 0.7$	0.99	9	5	50

Table 3.2: Classification accuracy and questionnaire length for synthetic data under the misspecified model. Results show the median questionnaire length, along with the 5th and 95-th percentiles, across different stopping rule conditions.

data partitions, with each fold serving as a test set. In contrast, the remaining nine folds provide training data for parameter estimation. Following established conventions in the VA literature, we adopt the missing-at-random assumption for handling incomplete questionnaire responses during model fitting. While potentially restrictive, this assumption represents current best practice and ensures comparability with existing methodological benchmarks (McCormick et al., 2016b; Kunihama et al., 2020b; Li et al., 2021).

Fixed-Length Questionnaire Performance

The empirical results in Figure 3.3 compare the performance of our adaptive questionnaire approaches with conventional static questioning protocols. Both the point estimate-based and posterior predictive scoring strategies achieve significantly better classification accuracy than traditional fixed-order questionnaires, with performance advantages becoming clear after just 10 questions. A notable finding emerges in the performance of our active designs: classification accuracy reaches optimal levels at approximately $30-40$ questions, followed by a slight decline in performance as the questionnaire length increases beyond this point. This relationship between questionnaire length and accuracy reflects limitations of our underlying analytical model, which uses simplified conditional independence assumptions that may not fully capture the complex symptom-cause relationships in real-world mortality data (Blanco et al., 2020; Li et al., 2020). The observed performance plateau and subsequent decline likely result from the accumulation of model specification errors as additional questions introduce noise rather than meaningful information under our simplified modeling framework. This phenomenon is well-documented in machine learning literature, where overly complex feature sets can reduce performance when the underlying model lacks sufficient sophistication to use the additional information effectively (Hastie et al., 2009).

While developing more sophisticated COD assignment algorithms represents an active research area (Wu et al., 2021a; Li et al., 2021), our findings provide clear evidence that substantial performance improvements are achievable even under simple modeling assumptions. Most importantly, the results demonstrate that our active questionnaire methodology enables highly accurate COD classification using only around 25% of the complete question set, representing a fourfold reduction in interview burden without compromising diagnostic accuracy.

Adaptive Stopping Rule Performance

We explore primary probability thresholds $p_{1\text{st}} \in \{0.7, 0.8, 0.9\}$ while maintaining a fixed discrimination parameter $d = 0.5$. Complete sensitivity analyses examining alternative d values are documented in Appendix B. We compare stopping rules based on point estimates against posterior predictive probability strategies with confidence levels $r \in \{0.5, 0.7\}$. Figure 3.4 shows the relationship between classification accuracy at questionnaire termination and median question requirements, organized by underlying COD. The results show that for most causes, classification accuracy does not change significantly with different threshold configurations. The main variation occurs in questionnaire length as stopping rules become stricter. This finding matches our earlier simulation observations, confirming that effective COD classification can be achieved with relatively few questions (James et al., 2011; Serina et al., 2015a). We notice that for specific causes, including maternal mortality, breast cancer, cervical cancer, drowning, homicide, and envenomation injuries from animals, the active questionnaire design achieves rapid convergence to high classification accuracy regardless of the stopping criterion used. This likely reflects the presence of highly distinctive symptom patterns that provide strong diagnostic signals for these particular causes (Chandramohan et al., 2021). The existence of such characteristic symptom clusters has significant practical implications, enabling interviewers to confidently terminate VA sessions much earlier when encountering these readily identifiable patterns of mortality. This cause-specific variation in diagnostic complexity suggests opportunities for developing tiered interview protocols that adapt termination criteria based on emerging diagnostic likelihood patterns (Shawon et al., 2021).

To place our methodology within the broader landscape of computational VA approaches, we benchmark our active questionnaire strategy against two established COD assignment algorithms: `InSilicoVA` (McCormick et al., 2016b) and `InterVA` (Byass et al., 2019). As shown in Figure 3.4, our active questionnaire design achieves classification accuracy that is comparable to, and often exceeds, that of both benchmark methods, despite

using substantially fewer questions. Our findings show that strategic question selection can potentially outperform comprehensive but inefficient data collection strategies. This suggests that current VA practices may be collecting substantial redundant information that contributes minimally to diagnostic accuracy while imposing an unnecessary burden on both respondents and data collection systems (Fottrell and Byass, 2010).

3.5 Conclusions and Discussions

This chapter presents a novel adaptive questionnaire design methodology specifically developed for VA applications. We introduce a Bayesian framework for estimating posterior predictive scores of questions based on KL information applied to question banks. Our approach accounts for comprehensive uncertainty quantification across any probabilistic COD assignment models, enabling both adaptive stopping rules and preservation of existing questionnaire flow structures. Empirical validation demonstrates performance improvements in COD classification accuracy across both fixed-length and adaptive-length questionnaire configurations.

The generalizability of our approach extends beyond VA applications to broader categories of medical and epidemiological surveys in resource-limited environments (Korenromp et al., 2003; Sankoh and Byass, 2014). The proposed active questionnaire design can be implemented for field deployment using the existing electronic data collection infrastructure that has become standard in VA programs (World Health Organization, 2022c). The system separates intensive model estimation, performed offline in centralized computing environments, from real-time question selection that requires only lightweight computations on standard tablet devices. Optimal tuning parameter selection for adaptive termination rules will require field validation studies, which are beyond the scope of this chapter.

Despite the promising results demonstrated in our work, we acknowledge several limitations. Our implementation employs relatively simple analytical models compared to state-of-the-art approaches in the current literature (Kunihama et al., 2020b; Li et al.,

2021; Wu et al., 2021a). While we observe meaningful advantages in posterior predictive scoring for termination decisions, the overall classification accuracy improvements over point estimate-based approaches remain modest. Our current framework does not fully utilize the extensive domain knowledge embedded in traditional VA instrument design, particularly regarding logical relationships and dependencies among questions. This domain knowledge could inform more sophisticated penalty function designs that better reflect clinical reasoning patterns and enhance guidance for selecting penalty parameters.

We note several important research directions that could address the methodological challenges of active questionnaire design for VA. First, extending our framework to accommodate ensemble modeling approaches represents a natural progression, given documented performance improvements in COD classification through model averaging techniques (Datta et al., 2021). Ensemble-based active designs could enhance robustness against model misspecification while providing more reliable uncertainty quantification for termination rules. Second, the growing literature on domain-adaptive VA algorithms highlights substantial heterogeneity in symptom-cause relationships across different populations and healthcare contexts (Li et al., 2021; Wu et al., 2021a). Our proposed adaptive design could be extended to explicitly account for population-specific variation through hierarchical modeling approaches or transfer learning methodologies. Third, understanding the human factors associated with adaptive questioning represents a critical research priority with direct implications for data quality and respondent welfare. A systematic investigation of interviewer and respondent responses to non-traditional question ordering would inform more practical implementations. Additionally, developing comprehensive cost-effectiveness models that consider question-specific costs, respondent burden, and accuracy trade-offs would provide useful guidance for deployment decisions (Thomas et al., 2018; Sanghera et al., 2018).

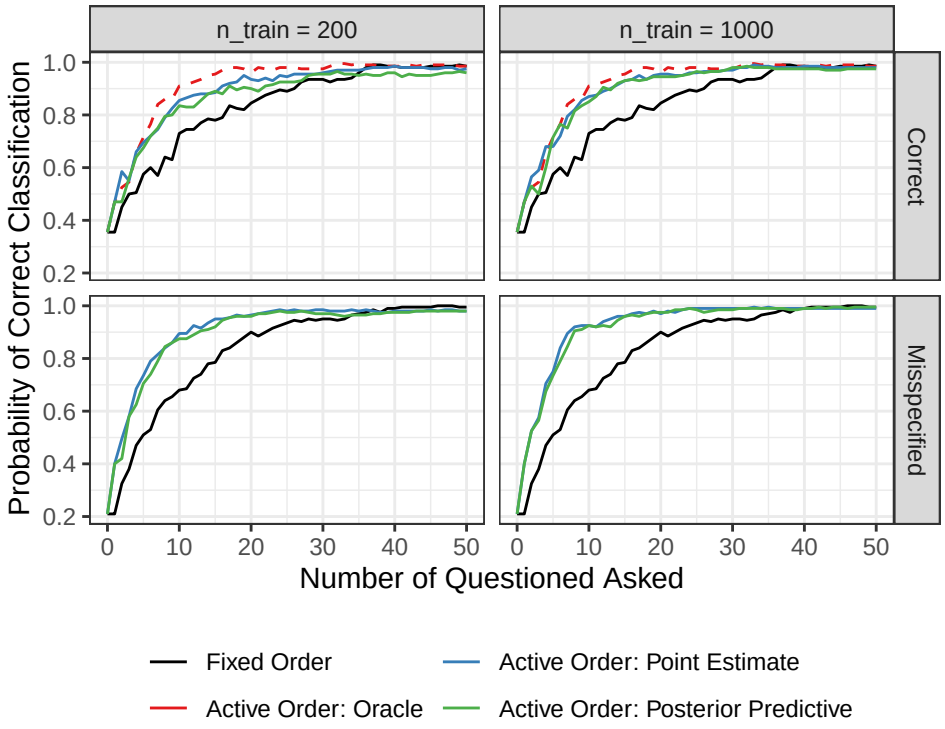


Figure 3.1: Classification accuracy across different questionnaire designs as the number of questions increases. Results are compared across four strategies for two training sample sizes (200 and 1000) and two scenarios (correct and misspecified models): fixed order (black), active order with point estimate scores (blue), active order with oracle parameters (red), and active order with posterior predictive scores (green). Active ordering methods consistently outperform fixed ordering, with the benefits most pronounced in smaller sample sizes.

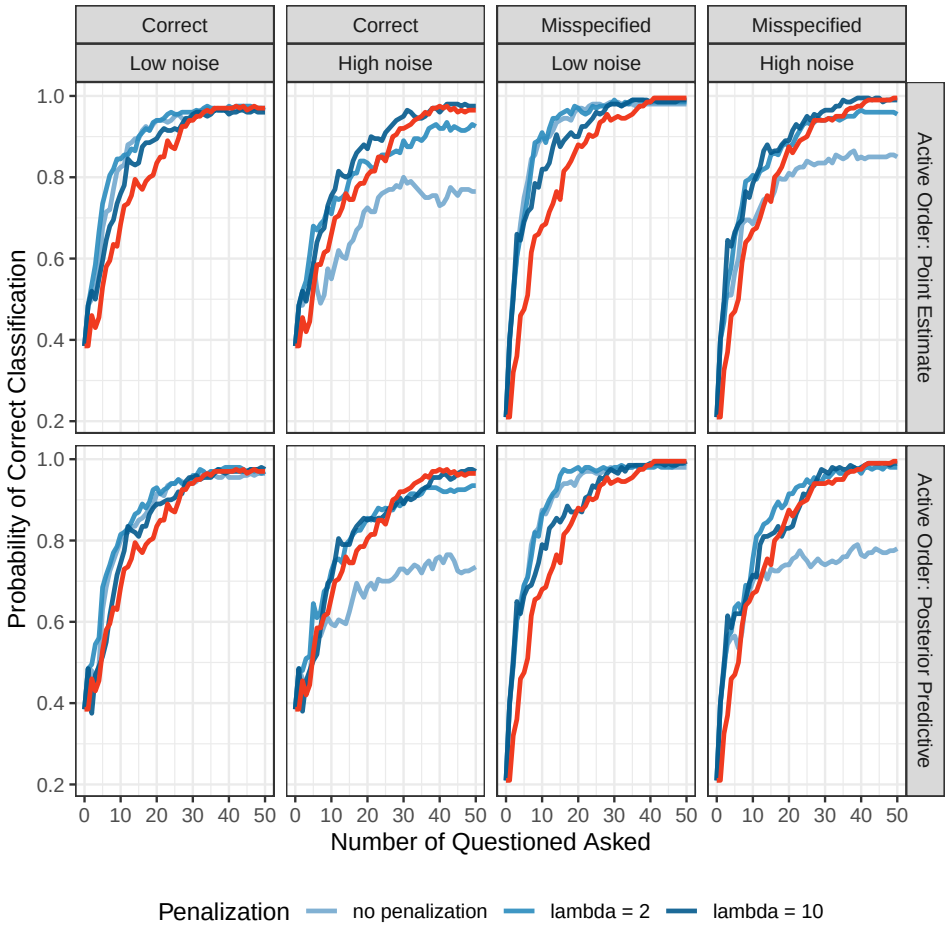


Figure 3.2: Classification accuracy of different questionnaire designs with order-induced response noise. Results are shown across four experimental conditions (correct/misspecified model with low/high noise levels) and two active ordering strategies (point estimate and posterior predictive). Three penalization levels are compared: no penalization (light blue), $\lambda = 2$ (medium blue), and $\lambda = 10$ (dark blue). The red curves represent fixed sequential ordering as a baseline comparison. Performance degrades with increased noise and model misspecification, with penalization helping to maintain accuracy in challenging conditions.

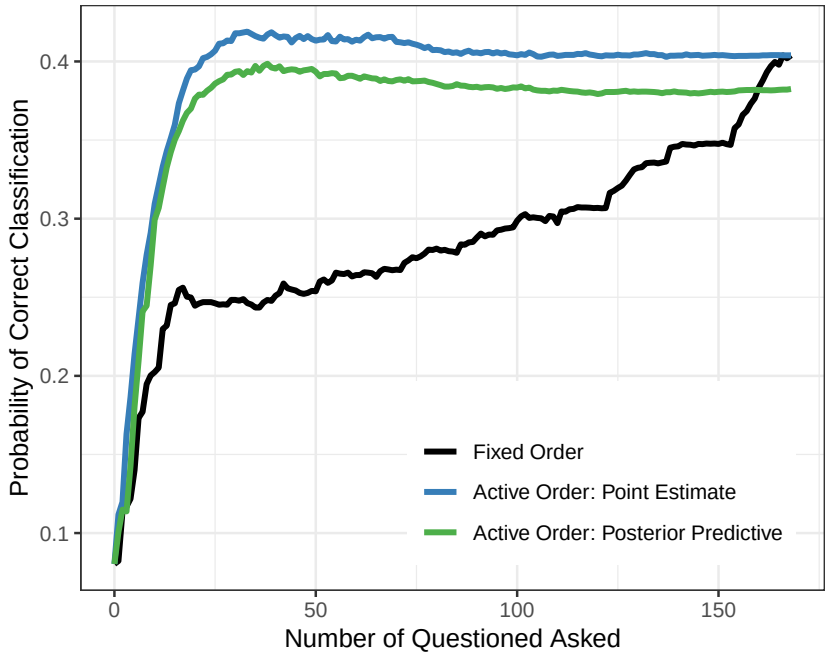


Figure 3.3: Classification accuracy of different questionnaire designs on PHMRC data as the number of questions increases. Three strategies are compared: fixed order (black), active order with point estimate scores (blue), and active order with posterior predictive scores (green). Active ordering methods achieve higher classification accuracy with fewer questions compared to fixed ordering, demonstrating substantial efficiency gains in the real-world VA context.

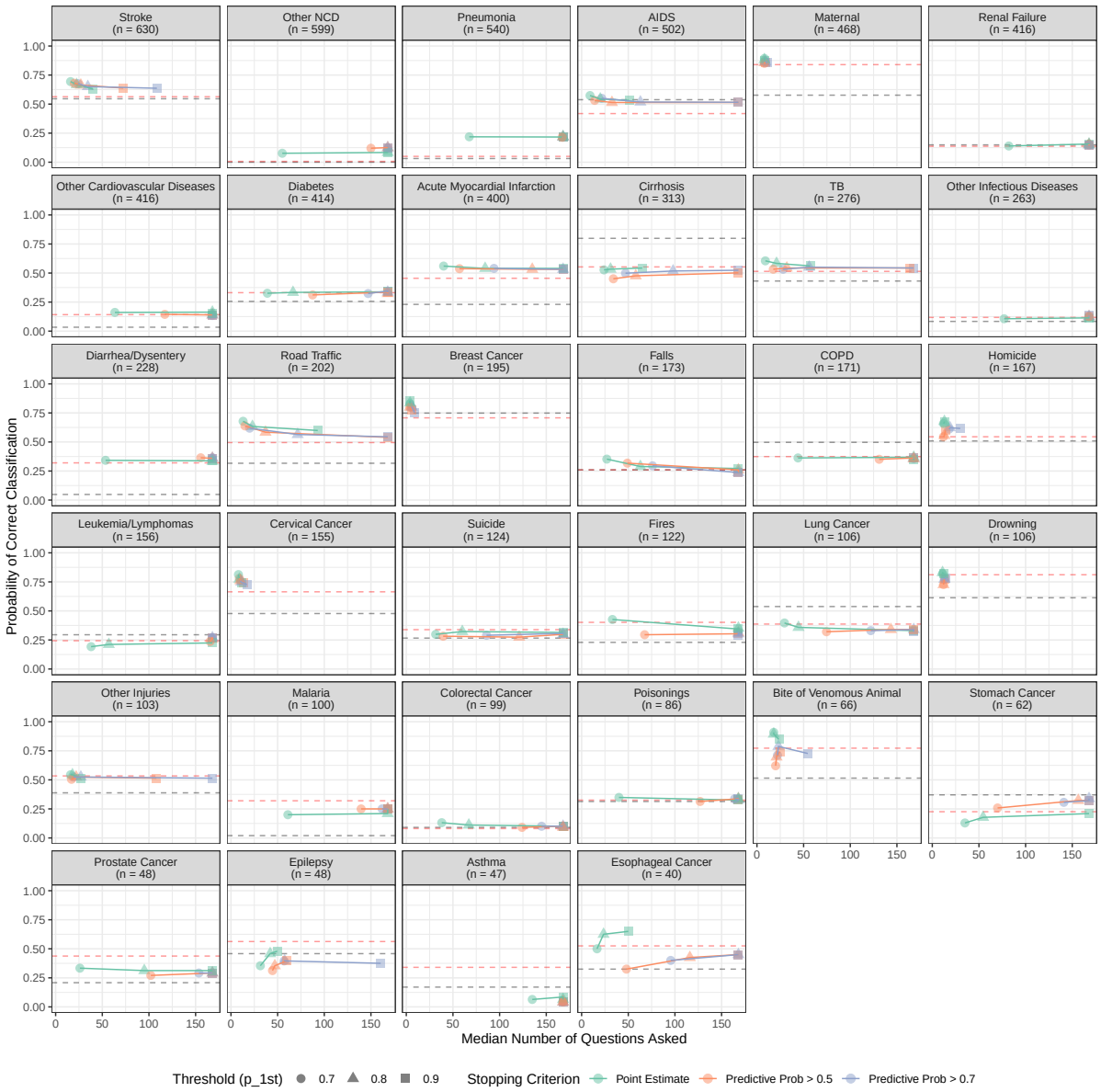


Figure 3.4: Proportion of correctly classified deaths by COD using different stopping criteria in PHMRC data. Each panel shows results for a specific cause ordered by sample size. Three stopping rules are compared: point estimate (green), pred $r > 0.5$ (orange), and pred $r > 0.7$ (purple), with different p_{1st} thresholds indicated by circle, triangle, and square symbols $(0.7, 0.8, 0.9$ respectively). Dotted horizontal lines represent accuracy benchmarks from established VA algorithms using all symptoms: `InSilicoVA` (red) and `InterVA` (blue).

Chapter 4

VALID INFERENCE USING LANGUAGE MODEL PREDICTIONS FROM VERBAL AUTOPSY NARRATIVES

4.1 Introduction

Verbal autopsy (VA) represents a critical epidemiological tool in public health research, particularly in settings with limited access to complete medical records (Soleman et al., 2006; Thomas et al., 2018). This approach determines cause-of-death (COD) by collecting information from surviving family members or caregivers who witnessed the circumstances surrounding a person's death in resource-constrained environments and regions with underdeveloped healthcare infrastructure (Baqui et al., 2006; Byass et al., 2012; Wahab et al., 2017). Contemporary VA interviews employ a dual-component structure, comprising structured questionnaires and unstructured narrative sections. The structured component captures specific symptoms and clinical indicators through standardized questions. In contrast, the narrative component allows respondents to provide contextual information about the illness and death circumstances in their own language (Chandramohan et al., 2021; Nichols et al., 2022).

Following data collection, COD assignment proceeds through clinical review by trained physicians or automated statistical algorithms that analyze symptom patterns. The trend toward algorithmic approaches reflects scalability advantages and reduced subjective variation (Murray et al., 2014; Serina et al., 2015a). Given financial and logistical barriers to comprehensive VA surveys, researchers use statistical modeling to extrapolate findings from representative samples to broader demographic patterns across age, sex, socioeconomic status, and geographic location.

This chapter presents a methodological framework for conducting valid statistical in-

ference using free-text VA narratives, as outlined in Figure 4.1. Our approach addresses two fundamental challenges that have limited the effective use of narrative data in mortality surveillance systems.

Our first contribution develops robust COD classification methods that rely exclusively on unstructured text narratives rather than structured questionnaires. Traditional VA interviews impose a substantial burden on both respondents and interviewers, typically requiring $60-90$ minutes to complete and involving emotionally challenging content that can worsen grief-related distress (Surek-Clark, 2020). Narrative-based approaches enable respondents to describe mortality circumstances using their own linguistic frameworks and cultural contexts, eliminating the need to navigate clinical terminology or respond to potentially irrelevant standardized questions. Building upon emerging research in natural language processing (NLP) applications for VA (Danso et al., 2013; Blanco et al., 2020; Manaka et al., 2022; Cejudo et al., 2023), our work incorporates large language models (LLM) and transformer-based architectures. It diverges from existing NLP studies in VA, which have primarily emphasized predictive accuracy as the primary evaluation metric (Naradowsky et al., 2012; Liu et al., 2019; Ye et al., 2021; Peskoff and Stewart, 2023). Instead, our framework prioritizes the downstream application of these predictions for valid statistical inference, recognizing that the ultimate goal of mortality surveillance systems is to support evidence-based decision-making in public health.

Our second contribution addresses the critical challenge of conducting valid statistical inference when COD labels are mainly predicted rather than clinically verified. The development and validation of COD prediction algorithms is challenging (Murray et al., 2011b; Fottrell and Byass, 2010; Murray et al., 2014; Clark et al., 2018b), primarily due to difficulties in constructing training datasets with reliably labeled CODs. While clinical review of VA cases could provide high-quality labels, this approach diverts scarce medical expertise from direct patient care, and traditional autopsy procedures remain infeasible in most settings where VA is employed. A natural approach to address limited labeled data involves aggregating VA datasets across multiple domains to create larger

training sets. However, this strategy introduces significant transportability issues due to heterogeneity in cultural practices, linguistic patterns, interview methodologies, and potentially underlying biological factors across different populations. Algorithms trained on data from one context frequently demonstrate degraded performance when applied to other domains.

To address these challenges, we develop a statistical inference framework that explicitly accounts for the additional uncertainty introduced when most COD labels are predicted rather than observed. We extend the prediction-powered inference (PPI) framework (`PPI++`; Angelopoulos et al., 2023a,b) to multinomial classification settings, demonstrating its broader applicability beyond binary and continuous outcome scenarios. We introduce `multiPPI++`, an estimator designed to enable valid statistical inference using predictions from arbitrary machine learning models, requiring only access to a small subset of ground truth labels for calibration purposes.

4.2 Population Health Metrics Research Consortium Data

Our empirical analysis uses the Population Health Metrics Research Consortium (PHMRC) dataset, established in 2005 to provide high-quality, standardized VA validation data with clinically verified COD determinations (Murray et al., 2011a). This dataset represents one of the few existing VA collections with comprehensive ground truth labels derived from traditional autopsy procedures, making it an invaluable resource for developing and benchmarking computational COD assignment algorithms (Serina et al., 2015a; James et al., 2011). The PHMRC data are collected across six study sites spanning four countries: *Andhra Pradesh* and *Uttar Pradesh* in India; *Bohol Island* in the Philippines; *Dar es Salaam* and *Pemba Island* in Tanzania; and *Mexico City* in Mexico. To illustrate the narrative data in PHMRC, consider this example: *"the deceased had been burnt and had lost mental balance and died within* 1.5 *hours of the accident"* (confirmed COD: *Fires*). This demonstrates how narrative responses capture critical contextual information about circumstances and timeline that structured questionnaires may miss.

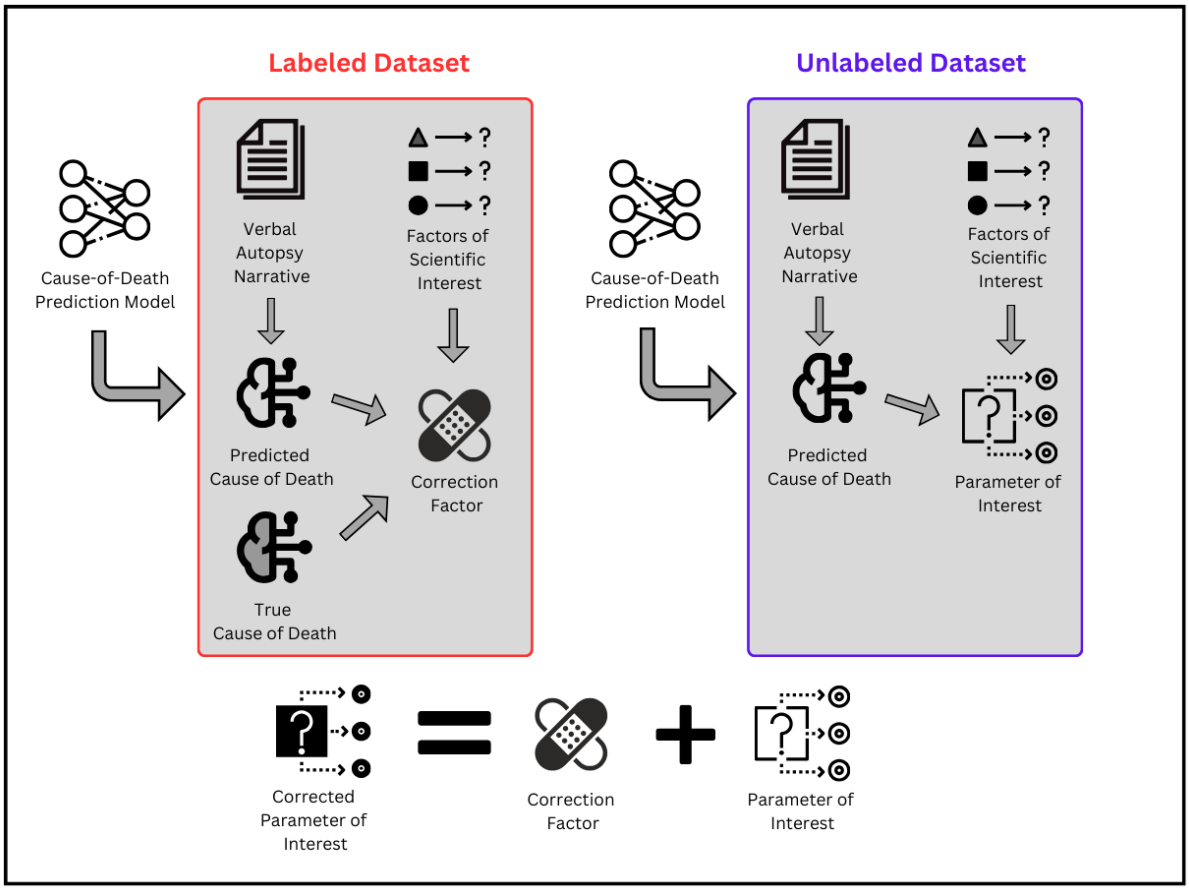


Figure 4.1: Overview of valid inference using `multiPPI++` for VA narratives. The method operates across two domains: a labeled dataset (red box), where ground-truth CODs are available, and an unlabeled dataset (blue box), where only predicted CODs are known.

We focus our analysis on adult mortality cases, excluding deaths in individuals under 12 years of age, using the de-identified narrative collection released by Flaxman et al. (2018). To facilitate meaningful statistical analysis while maintaining sufficient sample sizes, we aggregate the original COD classifications into five broad categories: non-communicable diseases, communicable diseases, external causes, maternal causes, and AIDS or tuberculosis (AIDS-TB). This categorization follows established International Classification of Diseases Tenth Revision (ICD-10) coding standards (McCormick et al., 2016b) and aligns with global burden of disease analytical approaches (Murray et al., 2020). Complete mapping details are provided in Appendix C.0.1.

Despite this consolidation, considerable heterogeneity persists in COD distributions across study sites, as demonstrated in Figure 4.2. While non-communicable diseases consistently represent the most prevalent cause category across all six sites, the second-ranking cause exhibits marked site-specific variation. Communicable diseases constitute the second most common COD in two sites, external causes predominate in three sites, and AIDS-TB represents the second-leading COD in one site. The magnitude of these distributional differences is equally noteworthy, with the ratio between the first and second most common COD varying considerably across sites. *Pemba Island* exhibits a relatively balanced distribution with an approximate $5 : 4$ ratio between non-communicable diseases and the second-leading COD. In contrast, *Mexico City* demonstrates pronounced dominance of non-communicable diseases with a $6 : 1$ ratio relative to the secondary COD.

These observations carry important implications for developing and deploying NLP algorithms for COD classification. While data from one study site may provide valuable insights for algorithm development in other locations, the substantial inter-site heterogeneity necessitates considerable caution when extrapolating model performance across diverse geographic and cultural contexts. This transportability issue represents a well-documented phenomenon in the VA literature (e.g., McCormick et al., 2016b), where algorithms trained on data from one population frequently demonstrate degraded perfor-

mance when applied to different settings. Without robust methodological approaches to address cross-population variation, implementing comprehensive COD surveillance systems would require establishing independent data collection and validation infrastructure at each new site, an endeavor that would prove prohibitively expensive.

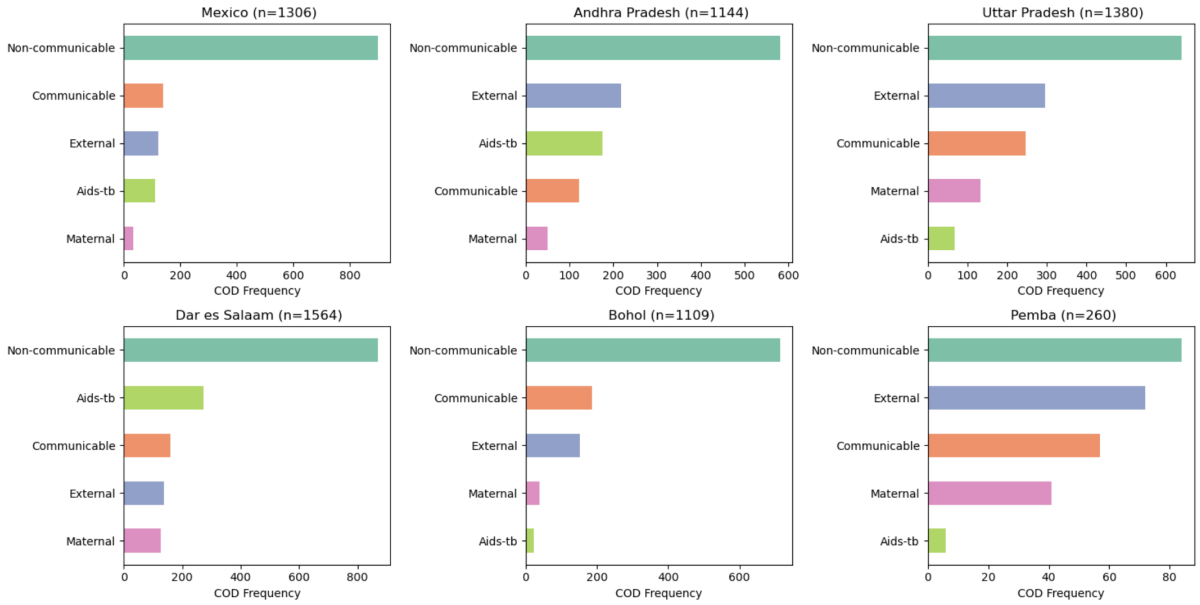


Figure 4.2: Frequency distribution of COD by study site. Non-communicable diseases represent the most prevalent cause across all six sites, though the relative distribution of cause categories varies substantially between locations.

4.3 Method

4.3.1 NLP for VA Narratives

The first objective of this chapter is to evaluate the predictive performance of contemporary NLP techniques for COD classification using exclusively free-text narrative data. This approach represents a shift from traditional VA analysis that relies primarily on structured questionnaire responses toward leveraging the rich contextual information embedded within narrative accounts of mortality circumstances. Our comparative analysis examines a comprehensive set of NLP techniques, ranging from classical statistical text

representation approaches to cutting-edge LLMs. We consider traditional bag-of-words (BoW) techniques that capture term frequency patterns, advanced pre-trained transformer architectures such as Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019), and contemporary LLMs, exemplified by the Generative Pre-trained Transformer (GPT) family (Lai et al., 2023).

For BoW representation generation, we employed `scikit-learn`'s `CountVectorizer`, `LabelEncoder`, and `TfidfVectorizer` (Kramer and Kramer, 2016; Fabian, 2011), complemented by `nltk`'s `word_tokenize` (Millstein, 2020) for text segmentation and normalization. Our preprocessing pipeline reduced the initial token count from 556713 to 218804 through systematic removal of low-frequency terms, stop words, and non-informative tokens. We evaluated three established supervised learning algorithms using the BoW representations paired with verified COD labels: Support Vector Machines (SVM), K-Nearest Neighbors (KNN), and Naive Bayes (NB) classifiers. The SVM employed a polynomial kernel of degree three with regularization parameter $C = 1.0$, selected based on cross-validation performance. Our KNN approach utilized cosine similarity metrics with $k = 9$ neighbors, determined through a systematic evaluation that began at $k = 3$ and incrementally increased until performance saturation was observed.

For transformer-based approaches, we configured a $\text{BERT}_{\text{BASE}}$ encoder with a five-layer architecture comprising dual input layers of dimension 256, a central transformer layer of dimension 768, an intermediate layer of dimension 512, and a task-specific output layer of dimension 5 corresponding to our COD categories. Model training proceeded over two epochs to prevent overfitting while ensuring adequate parameter convergence.

Given the de-identified nature of our narrative data, we evaluated OpenAI's LLMs, including GPT-3.5, GPT-3.5-turbo, GPT-4, and GPT-4-32K variants (Biswas, 2023). Following preliminary comparative analysis, we selected GPT-4-32K for primary evaluation due to its expanded context window capacity, which enables processing of longer narrative texts without truncation. All LLMs employed zero-shot prompting strategies detailed in Figure 4.3 with temperature parameters set to zero to ensure deterministic

and reproducible outputs. Across all model options, we maintained a consistent experimental design through 80/20 train-test data splitting, with COD labels withheld from test sets to evaluate predictive accuracy and F1-score metrics.

Prompt Engineering for LLMs

The development of prompting strategies for LLMs presented several challenges, particularly generating constrained outputs that align precisely with our predefined COD labels: communicable, non-communicable, external, maternal, and AIDS-TB. Initial prompt tuning frequently encountered semantically appropriate but structurally inconsistent responses, such as "the patient died due to external causes", which, while conceptually correct, required complex post-processing procedures including regular expression matching and fuzzy string algorithms for systematic validation. We address this issue by implementing structured HTML-style tagging within the prompt architecture. As shown in Figure 4.3, the prompt design incorporates several key elements: explicit narrative tagging through structured markup, minimal contextual descriptions for each COD category to provide semantic grounding, comprehensive enumeration of acceptable output options to constrain response space, and direct procedural instructions emphasizing adherence to the `<narrative><label><option>` tagging framework. This approach significantly improved output consistency.

Future refinements may incorporate few-shot prompting approaches with carefully selected domain-specific exemplars, or retrieval-augmented generation (RAG) frameworks that leverage larger narrative databases to provide contextual grounding for classification decisions. Such enhancements could potentially improve both accuracy and consistency while maintaining the efficiency advantages of our current zero-shot approach; however, they are beyond the scope of this chapter.

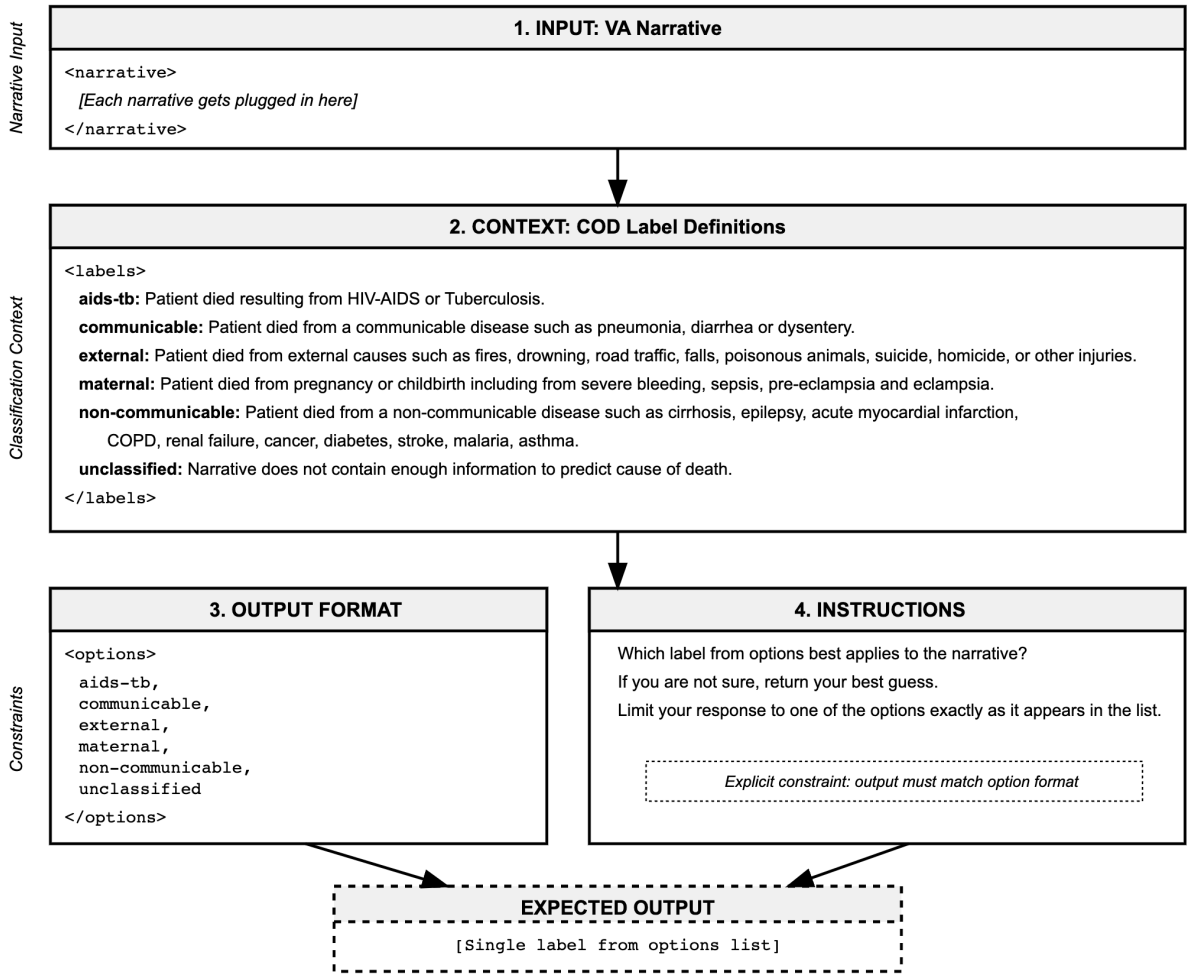


Figure 4.3: GPT-4 zero-shot prompt structure for COD prediction from VA narratives. The prompt framework consists of four components: (1) narrative input where individual VA texts are inserted, (2) classification context providing definitions for six COD categories (AIDS-TB, communicable, external, maternal, non-communicable, and unclassified), (3) structured output format constraining responses to the predefined options, and (4) explicit instructions directing the model to select the most appropriate label. The system generates a single COD classification from the available options for each narrative input.

4.3.2 Valid Statistical Inference with `multiPPI++`

The second objective of this chapter is to address the statistical challenge of conducting valid inference on covariate relationships when COD labels are mostly predicted rather than clinically verified. The inherent imperfection of NLP models introduces systematic bias into downstream statistical analyses, necessitating principled correction methods before reliable conclusions can be drawn from such data. While the PHMRC dataset provides comprehensive ground truth labels derived from traditional autopsy procedures, this level of verification is impractical for most global health surveillance applications. In resource-constrained settings, COD determination relies exclusively on VA data without access to verification means. Emerging research has shown that using algorithmically-derived outcomes for downstream statistical inference can lead to biased coefficient estimates and anti-conservative confidence intervals, compromising the reliability of subsequent public health decision-making (Wang et al., 2020; Angelopoulos et al., 2023a,b; Miao et al., 2023; Egami et al., 2023; Ogburn et al., 2021; Hoffman et al., 2024). This underscores the practical importance of developing statistical frameworks that can accommodate predicted outcomes while maintaining inferential validity.

Theoretical Framework

To address these inferential challenges, we employ and extend the recently developed PPI approach (Angelopoulos et al., 2023a) and its enhanced variant `PPI++` (Angelopoulos et al., 2023b), adapting these frameworks for multinomial classification contexts. These methods operate by leveraging small subsets of verified ground-truth labels to systematically correct coefficient estimates and standard errors derived from regressions that utilize predicted outcomes.

Our data structure comprises n labeled observations

$$\mathcal{L} = \{(Y_{l,i}, \hat{Y}^f_{l,i}, X_{l,i}, Z_{l,i}); i = 1, \ldots, n\},$$

and N unlabeled observations

$$\mathcal{U} = \{(\hat{Y}^f_{u,i}, X_{u,i}, Z_{u,i}); i = n+1, \ldots, n+N\}.$$

Here, Y denotes the clinically verified COD, $\hat{Y}^f$ denotes the predicted COD from the NLP model $f(Z)$ applied to narrative text Z, and X captures covariates of analytical interest such as demographic characteristics, geographic indicators, or temporal variables.

In the statistical inference paradigm, parameter estimation of θ can be formalized as an optimization problem minimizing an objective function:

$$\theta^{\text{Ground Truth}} = \arg\min_{\theta \in \mathbb{R}^d} \mathbb{E}[l_\theta(X, Y)],$$

where $l_\theta : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ represents a convex loss function parameterized by $\theta \in \mathbb{R}^d$. In our context, $l_\theta(X, Y)$ corresponds to the negative log conditional likelihood of multinomial logistic regression.

In practice, substituting the verified COD Y with predicted values $\hat{Y}^f$ yields a naive estimator:

$$\theta^{\text{Naive}} = \arg\min_{\theta \in \mathbb{R}^d} \mathbb{E}[l_\theta(X, \hat{Y}^f)].$$

This naive estimation assumes near-perfect correspondence between predicted and true COD labels. This assumption is frequently violated when algorithms are applied across diverse populations or contexts where the training data may not adequately represent the target domain.

Both `PPI++` and our extension `multiPPI++` seek to optimally leverage information from both labeled and unlabeled data sources through the following rectified objective function:

$$\theta^{\text{PPI++}}_\lambda = \arg\min_{\theta \in \mathbb{R}^d} \left\{ \mathbb{E}[l_\theta(X_l, Y_l)] + \lambda \left(\mathbb{E}[l_\theta(X_u, \hat{Y}^f_u)] - \mathbb{E}[l_\theta(X_l, \hat{Y}^f_l)] \right) \right\},$$

where $\lambda \in [0, 1]$ serves as a data-adaptive tuning parameter that optimally balances the utilization of verified versus predicted labels. This formulation combines the ground-truth estimator derived from the labeled subset with the additional sample size and

information provided by the naive estimator, incorporating the unlabeled subset. The key insight underlying this approach lies in the rectification term $\mathbb{E}[l_\theta(X_u, \hat{Y}_u^f)] - \mathbb{E}[l_\theta(X_l, \hat{Y}_l^f)]$, which quantifies the systematic bias introduced by substituting predicted for true labels.

We extend the `PPI++` approach from its original binary and continuous outcome contexts to multinomial classification settings. While Angelopoulos et al. (2023b) acknowledge the theoretical feasibility of such extensions, they provide neither detailed derivations nor address the substantial overparameterization challenges that arise in multiclass contexts. Our `multiPPI++` variant employs an alternative parameterization that circumvents identifiability issues while completing the theoretical development for multinomial logistic regression correction. For multinomial classification problems involving K classes with outcomes $Y_i \in \{0, 1, \ldots, K-1\}$, we specify the loss function as:

$$l_\theta(X, Y) = -\frac{1}{n} \sum_{i=1}^{n} X_i^T \theta_{Y_i} + \log \left(\sum_{k=1}^{K-1} e^{X_i^T \theta_k} \right),$$

where one reference category is excluded in parameterization to ensure model identifiability. We provide theoretical derivations and asymptotic properties in Appendices C.0.2 and C.0.3 with the complete `multiPPI++` algorithmic framework detailed in Algorithm 1.

4.4 Experiments

4.4.1 Experimental Setup

To assess the performance of our proposed framework, we designed an experimental paradigm that simulates realistic deployment scenarios where ground-truth COD is unavailable. We systematically removed ground-truth COD labels from the PHMRC dataset. We replaced them with predictions generated through the NLP methods detailed in Section 4.3.1. We employ a leave-one-site-out cross-validation protocol. For each of the six PHMRC study sites, we train all the NLP models exclusively on narrative data from the five remaining sites and subsequently apply these trained models to predict COD labels

for the withheld site, as illustrated in Figure 4.4. This approach directly addresses the transportability challenges inherent in cross-population deployment of mortality surveillance algorithms, providing realistic performance evaluation when models encounter populations absent from their training distributions (Botsis et al., 2010; Korenromp et al., 2003).

Our evaluation has two objectives: predictive performance assessment through comparison of predicted outcomes against withheld ground-truth COD labels using standard classification metrics (accuracy and F1-scores), and inferential validity evaluation through downstream statistical modeling. The latter involves implementing an epidemiological analysis wherein site-specific COD distributions are regressed against decedent age, serving as a prototypical example of analytical tasks commonly performed using VA data in public health practice (Clark et al., 2018a; Thomas et al., 2018).

To construct the mixed labeled-unlabeled data structure required for `multiPPI++` implementation, we retain 20% of ground-truth COD as labeled observations while substituting the remaining 80% with NLP-generated predictions. This allocation reflects realistic resource constraints in settings where comprehensive clinical verification remains prohibitively expensive while ensuring sufficient labeled data for bias correction (Sankoh and Byass, 2014). We compared the proposed `multiPPI++` estimator with two alternatives: the ground-truth estimator, which utilizes complete, verified COD labels (representing the theoretical performance ceiling under perfect information), and the naive estimator, which treats all predicted labels as directly observed (reflecting current uncorrected practice settings). While the ground-truth estimator is typically infeasible in practical settings, its inclusion provides essential context for interpreting the magnitude of performance gains achieved through our proposed approach. Given the inter-site heterogeneity in COD distributions and the intentional exclusion of target sites from model training, we anticipate systematic disparities between regression coefficients estimated using predicted versus ground-truth COD labels. These differences reflect inherent cross-population transportability issues and specific prediction errors introduced by each NLP

approach.

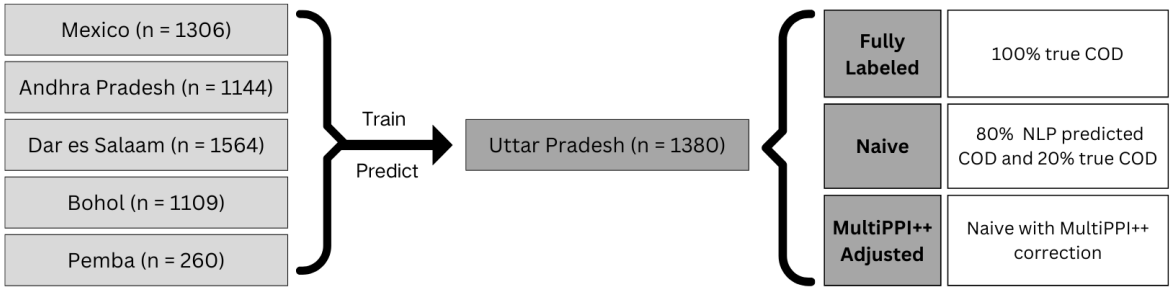


Figure 4.4: Leave-one-out synthetic VA transportability experiment design. The study employs a cross-site validation approach where machine learning models (classic and BERT) are trained on VA narratives from five sites (*Mexico, Andhra Pradesh, Dar es Salaam, Bohol, and Pemba*) and applied to predict COD in the held-out *Uttar Pradesh* dataset. A zero-shot GPT-4 model provides predictions without site-specific fine-tuning. Three analytical approaches are compared using the *Uttar Pradesh* predictions: (i) fully labeled analysis using 100% true COD, (ii) naive analysis combining 80% model-predicted and 20% true COD, and (iii) `multiPPI++` adjusted analysis that applies bias correction to the naive approach.

4.4.2 Predictive Performance

Implementing the leave-one-site-out cross-validation procedure described in Figure 4.4, we trained each NLP model on narrative data from five study sites while reserving the sixth site for assessment. As evidenced by the heterogeneous patterns illustrated in Figure 4.2, certain COD categories dramatically outnumber others both within individual sites and across the broader dataset. We address this through comprehensive sensitivity analyses, documented in Appendix C.0.4, which employ stratified sampling approaches. These analyses utilized balanced 80/20 splits across all COD categories, ensuring proportional representation during model training phases. These stratified approaches yielded results that were not substantively different from those obtained in our primary analy-

ses using natural class distributions, demonstrating the robustness of our methodological framework to realistic data imbalances encountered in VA studies. Our findings indicate that `multiPPI++` maintains inferential validity even when utilizing predictions from models trained on naturally imbalanced datasets. This robustness represents a significant practical advantage, eliminating the need for complex data preprocessing procedures while preserving authentic mortality distribution patterns characteristic of real-world populations.

We provide a detailed presentation of results from *Uttar Pradesh* as a representative case study, while comprehensive results from all sites are presented in Appendix C.0.4. Performance metrics for the highest-performing models are summarized in Table 4.1 and visualized through confusion matrices in Figure 4.5. All evaluated models achieved accuracy scores ranging from 0.60 to 0.75, with GPT-4 achieving the highest accuracy of 0.75. F1-scores demonstrated greater convergence across approaches. All methods except KNN achieved F1-scores between $0.71-0.74$, with KNN achieving a lower F1-score of 0.66. This pattern suggests that while overall classification performance remains relatively consistent across methodologies, GPT-4's performance reflects its ability to handle the nuanced linguistic patterns and contextual relationships embedded within VA narratives. GPT-4 achieved performance levels that match or exceed the best-reported results in the established VA literature (James et al., 2011; Murray et al., 2014; Serina et al., 2015b), despite using only free-text narratives without access to the structured questionnaire data typically employed in such analyses. This finding supports our hypothesis that these underutilized data sources contain sufficient signal for reliable COD classification when processed through appropriate methods.

A unique feature of LLMs relative to conventional machine learning approaches is their capacity to generate output classifications that extend beyond predefined taxonomic boundaries. This proved unexpectedly valuable for data quality assessment in our analysis. GPT-4 classified 1503 out of 6763 narratives (22.2%) as "*unclassified*", reflecting the model's recognition that these texts contained insufficient diagnostic information

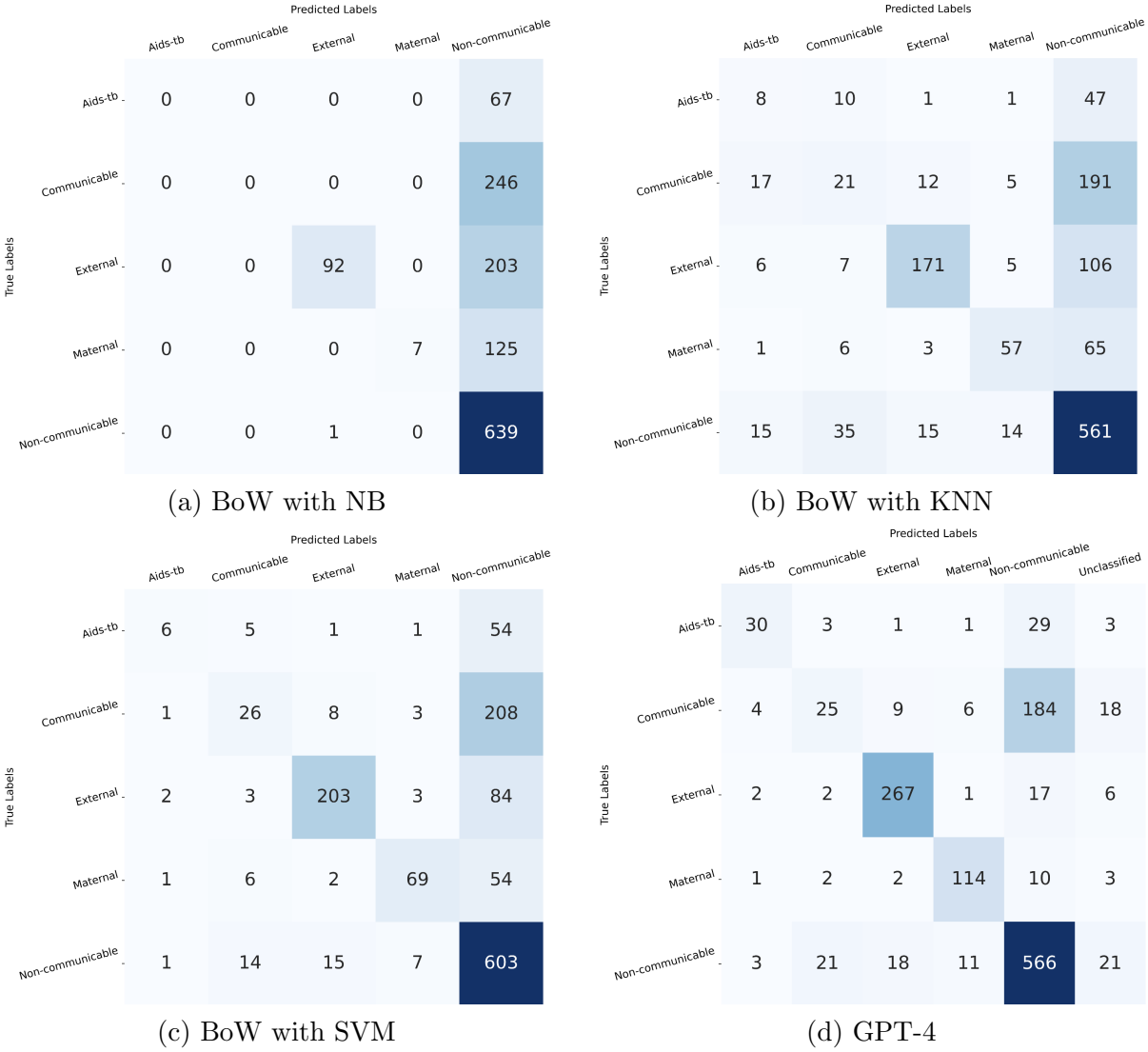
Metrics	BoW w/ SVM	BoW w/ KNN	BoW w/ NB	BERT	GPT-4
Accuracy	0.68	0.63	0.60	0.55	0.75
F1-Score	0.74	0.66	0.72	0.71	0.73

Table 4.1: Comparative performance metrics across NLP methodologies for *Uttar Pradesh* validation site. GPT-4 achieves the highest accuracy and competitive F1-score, outperforming traditional BoW approaches with SVM, KNN, and NB, as well as BERT.

for reliable COD determination. Review of these unclassified cases confirmed their lack of meaningful diagnostic content, with the modal "*unclassified*" narrative being "*the interviewee thanked the interviewer*" ($n = 178$). Additional frequently occurring uninformative content included procedural acknowledgments, incomplete responses, and transcription artifacts that provided no clinical insights. This pattern indicates that GPT-4 can effectively identify systematic data quality issues within the PHMRC dataset that had escaped conventional preprocessing procedures. While surface-level accuracy comparisons might suggest roughly equivalent performance across methods, GPT-4's ability to appropriately abstain from classification when presented with uninformative content suggests that conventional accuracy metrics for other approaches may be artificially inflated due to forced classification of diagnostically meaningless narratives. This finding has important implications for operational deployment, as it provides a built-in mechanism for identifying cases requiring additional data collection or alternative analytical approaches (Rajkomar et al., 2018; Chen et al., 2021).

4.4.3 Inferential Validity

Following prediction generation across all site-model combinations, we propagated these results through our downstream inferential framework to evaluate the practical utility of `multiPPI++`. For focused presentation, we highlight results from the four best-performing models for the *Uttar Pradesh* site (NB, KNN, SVM, and GPT-4), with detailed re-



(a) BoW with NB (b) BoW with KNN

(c) BoW with SVM (d) GPT-4

Figure 4.5: Confusion matrices for COD classification performance across NLP methodologies using *Uttar Pradesh* validation data. The matrices compare four approaches: (a) BoW with NB, (b) BoW with KNN, (c) BoW with SVM, and (d) GPT-4. The dominant misclassification pattern involves assignment to the non-communicable disease category, reflecting both the high baseline prevalence of this cause and systematic challenges in distinguishing between related mortality classifications. NB exhibits the most severe bias toward non-communicable predictions, while GPT-4 demonstrates more balanced classification performance across CODs.

sults presented in Figure 4.6, displaying point estimates and 95% confidence intervals for ground-truth, naive, and `multiPPI++` corrected inference results. For models exhibiting relatively lower F1-scores, specifically NB and SVM, we observed large discrepancies between ground truth and prediction-based regression coefficients, demonstrating the profound impact of substituting 80% of the ground-truth COD labels with algorithmic predictions on downstream statistical inference. The `multiPPI++` correction successfully rectifies these biased estimates, restoring coefficients to levels closely approximating the ground-truth baseline.

A particularly intriguing finding emerges from examination of KNN and GPT-4 performance patterns. Despite KNN achieving the lowest F1-score among evaluated methods, the naive regression coefficients derived from KNN predictions demonstrated closer alignment with ground-truth values compared to nominally superior-performing approaches. Similarly, GPT-4-derived coefficients exhibited reduced deviation from baseline despite its intermediate F1-score performance. This apparent paradox suggests complex, non-linear relationships between conventional predictive performance metrics and downstream inferential quality. Figure 4.7 illustrates the lack of association between F1-score and the `multiPPI++` tuning parameter λ, providing additional evidence that traditional classification metrics may not reliably predict inferential utility. This finding has important implications for model selection in applied settings, suggesting that optimization for predictive accuracy alone may not yield optimal outcomes for downstream epidemiological analysis. Instead, practitioners should consider the specific analytical objectives and covariate relationships of interest when selecting models for inferential applications (Mullainathan and Spiess, 2017).

Examination of confidence interval widths reveals insights regarding the relative precision of different inferential approaches. As expected, ground truth estimates exhibit the narrowest confidence intervals, reflecting both the larger effective sample size and the elimination of prediction-induced uncertainty. Conversely, naive estimates and `multiPPI++` corrected results demonstrate remarkably similar interval widths, suggesting that the bias

correction procedure does not substantially increase statistical uncertainty. This finding is theoretically noteworthy given that `multiPPI++` confidence intervals mathematically represent the sum of naive estimation uncertainty plus correction term variance. The observed similarity in interval widths suggests that the correction term contributes minimal additional uncertainty, indicating that transportability bias can be effectively addressed without incurring a substantial information cost. This efficiency represents a significant practical advantage, enabling bias correction while preserving statistical power for hypothesis testing and effect size estimation (Gelman et al., 2014; Manski, 2009). The apparent independence between the adaptive tuning parameter λ and prediction accuracy provides additional insight into the mechanism underlying the performance of `multiPPI++`. Since λ represents an optimal weighting function determined by both outcome distributions and covariate relationships, its independence from accuracy metrics suggests that superior predictive performance does not necessarily translate to enhanced parameter estimation.

Figure 4.6 demonstrates that `multiPPI++` improves statistical estimate accuracy through a post-hoc correction term estimated using small amounts of labeled data. This correction operates independently of the prediction generation process, rectifying biased statistical inference parameters and calibrating uncertainty intervals that would otherwise be deflated due to substituting predicted for verified outcomes. This distinction is important because VA analysis aims for reliable population-level inference about covariate-outcome relationships rather than just accurate individual case classification. The parameters quantifying associations between demographic factors and COD distributions represent the primary outputs that support evidence-based public health interventions. By ensuring the validity of these inferential targets, `multiPPI++` enables confident utilization of NLP predictions for epidemiological analysis without compromising scientific rigor. This approach extends to other domains where NLP predictions substitute for human-verified labels in statistical inference applications, including sentiment analysis for policy evaluation, clinical note processing for health outcomes research, and content classification for

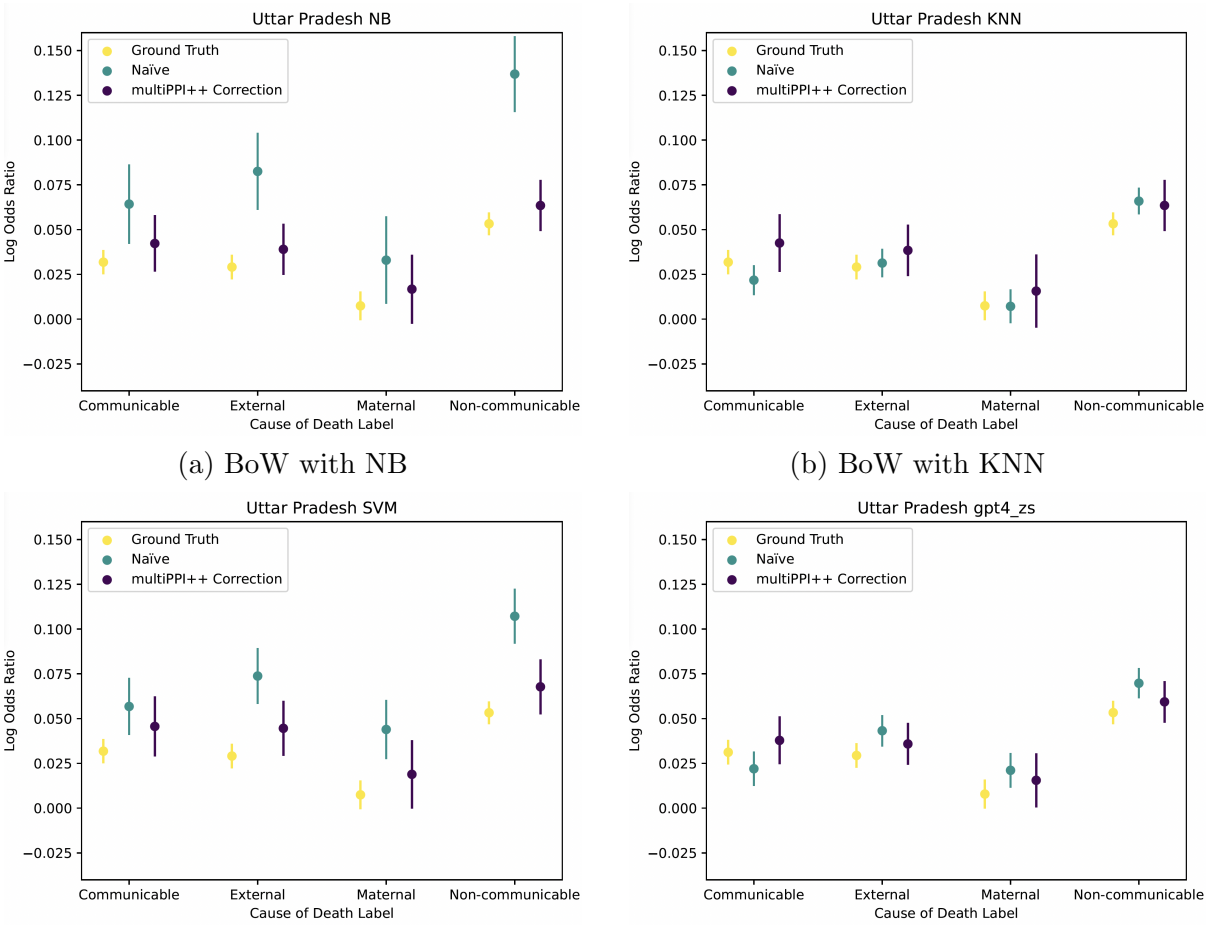
social science investigations (Egami et al., 2022; Knox et al., 2022; Grimmer and Stewart, 2013).

4.5 Conclusions and Discussions

This study presents NLP approaches for predicting COD from VA narratives and develops a statistical method for valid inference using these predictions. The research emphasizes the valid interpretation of parameters for evidence-based decision-making in diverse health system contexts (D'Ambruoso et al., 2017; Fottrell and Byass, 2010). Integrating narrative-based COD classification with bias-corrected statistical inference represents a shift toward utilizing underused data sources while maintaining scientific rigor for public health applications. Our findings demonstrate that computational approaches can extract meaningful information from free-text mortality descriptions, potentially improving the scalability and efficiency of global mortality surveillance systems (Thomas et al., 2018; Chandramohan et al., 2021).

Our empirical analysis reveals counterintuitive findings that challenge assumptions about predictive performance and inferential validity. Superior classification accuracy in COD predictions does not necessarily improve parameter estimation quality for statistical inference within the examined accuracy ranges. This disconnect has significant implications for model selection in public health contexts, where reliable parameter estimation typically takes precedence over individual case classification (Mullainathan and Spiess, 2017). When researchers prioritize accurate parameter estimation for epidemiological analysis, models optimized for predictive accuracy may not be optimal for statistical inference. Our findings suggest that computationally inexpensive approaches, such as KNN, offer inferential utility comparable to that of sophisticated LLMs like GPT-4, despite substantial differences in computational complexity and cost.

The economic implications are substantial. Implementing BoW representations with classical machine learning algorithms (KNN, NB, and SVM) incurred negligible computational costs while achieving comparable inferential performance to GPT-4-based



(a) BoW with NB

(b) BoW with KNN

(c) BoW with SVM

(d) GPT-4-32k

Figure 4.6: `multiPPI++` correction of log odds ratio estimates for multinomial regression coefficients by age in *Uttar Pradesh* across four NLP models. The analysis compares ground truth estimates (yellow), naive estimates using predicted COD (blue), and `multiPPI++` corrected estimates (black) for four categories of COD. Results demonstrate that `multiPPI++` correction effectively recovers point estimates closer to ground truth values while appropriately reflecting increased uncertainty through wider confidence intervals. This pattern holds consistently across (a) BoW with NB, (b) BoW with KNN, (c) BoW with SVM, and (d) GPT-4.

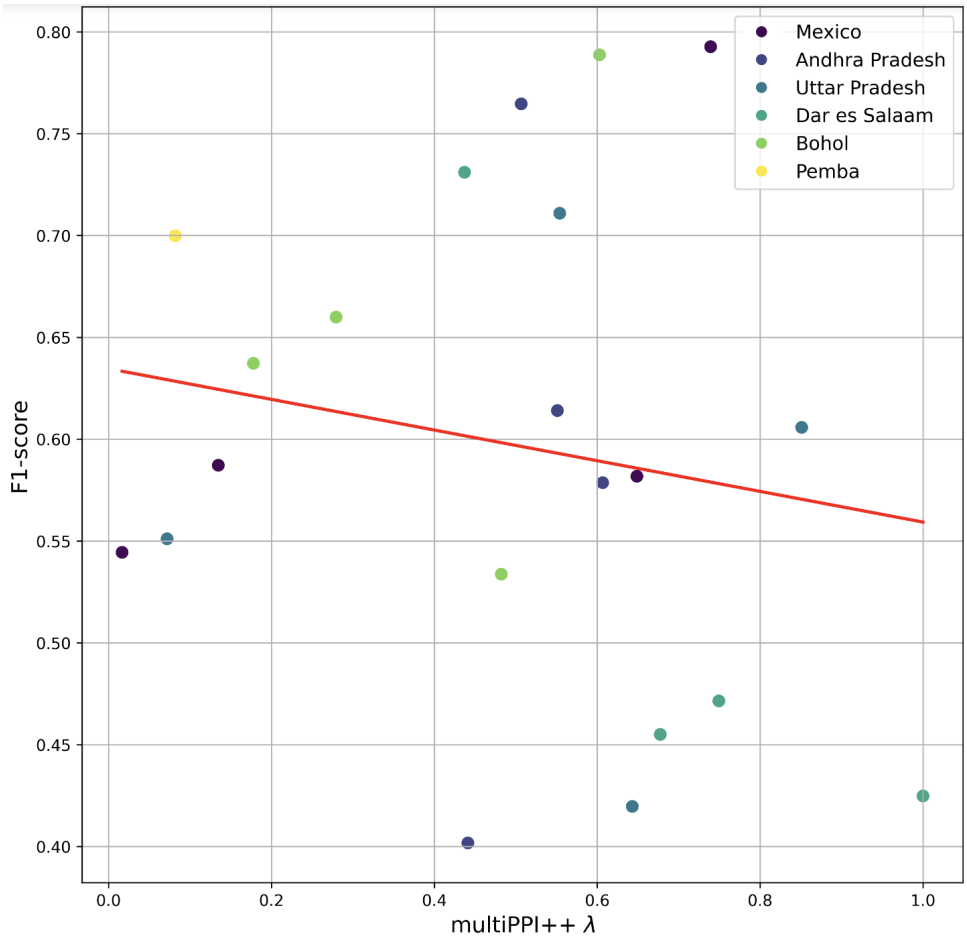


Figure 4.7: Relationship between F1-scores and `multiPPI++` λ values across study sites. The scatter plot shows F1-score performance plotted against `multiPPI++` λ parameters for six sites (*Mexico, Andhra Pradesh, Uttar Pradesh, Dar es Salaam, Bohol, and Pemba*), with a fitted trend line indicating a slight negative association. While no strong global relationship emerges between `multiPPI++` λ and F1-score across all sites, distinct patterns become apparent when examining individual sites, suggesting potential site-specific effects that may reflect a Simpson's paradox phenomenon where the overall trend masks more nuanced relationships within subgroups.

approaches that required approximately $3000 USD in computational expenses. This cost-performance relationship suggests that resource-constrained public health programs can achieve their analytical objectives through efficient modeling approaches, reserving expensive computational resources for scenarios where additional investment demonstrably improves inferential outcomes (Gibbons et al., 2017).

Class imbalance in COD distributions significantly influences predictive performance and inferential validity across all evaluated methodologies. While direct balancing across COD categories remains impractical given authentic epidemiological patterns, this challenge raises fundamental design questions for research planning and resource allocation. Practitioners must consider optimal allocation of limited labeling budgets to maximize inferential validity. Strategic oversampling of rare but epidemiologically important causes might enhance statistical power for detecting meaningful associations while preserving natural distributional characteristics essential for population-level inference (King and Zeng, 2001; Weiss and Provost, 2003). Adaptive sampling strategies that iterate between model development and targeted validation could optimize the trade-off between labeling costs and inferential precision.

Our analysis employed a five-category taxonomic framework that may not optimally capture information content within VA narratives. Alternative categorization schemes informed by clinical expertise, data-driven clustering, or hierarchical COD frameworks might provide richer information for predictive modeling and inference. Investigating optimal taxonomic granularity represents an important avenue for enhancing narrative-based mortality surveillance systems (Murray et al., 2020; Roth et al., 2018).

Another limitation stems from our focus on English-language narratives, which inadequately represent linguistic diversity in global VA programs. NLP techniques demonstrate degraded performance when applied to non-English languages (Navigli et al., 2023; Lai et al., 2023; Hirschberg and Manning, 2015). Since VA programs operate predominantly in multilingual, low-resource settings, this performance degradation represents a critical concern for equitable global health surveillance. Linguistic bias in automated COD

classification could disadvantage populations whose languages are underrepresented in training datasets, exacerbating health inequities and compromising cross-national mortality comparisons (Surek-Clark, 2020; Byass et al., 2012). Our simulation study, which uses exclusively English-language PHMRC narratives, does not account for the differential translation quality and cultural context encountered in multilingual implementations.

Future research should prioritize multilingual NLP models pre-trained on diverse linguistic corpora and medical terminology (Peskoff and Stewart, 2023; Singhal et al., 2023). Investigating culturally adapted taxonomic frameworks and validation approaches that account for population-specific expressions of illness and mortality could enhance both the accuracy and cultural appropriateness of automated VA systems. Such developments are essential for ensuring that computational approaches contribute to rather than exacerbate global health disparities (Kleinberg et al., 2015; Barocas and Selbst, 2016).

Chapter 5

BAYESIAN FRAMEWORK FOR MORTALITY ESTIMATION WITH INCOMPLETE AGE CLASSIFICATIONS

5.1 Introduction

Estimating under-five child mortality requires precise age stratification to capture heterogeneous risk profiles across different developmental stages. Mortality risk and cause-specific patterns vary dramatically across the under-five period, with the neonatal period ($0 - 28$ days) representing the highest-risk window, accounting for 48% of all under-five deaths globally despite comprising only 5.5% of the under-five age period (Mulick et al., 2022). Within this critical window, approximately 75% of neonatal deaths occur in the first week of life, with nearly one million newborns dying within 24 hours of birth (Zamboni et al., 2021; Sepanlou et al., 2022). Birth complications and neonatal disorders concentrate almost exclusively in the first month of life. At the same time, infectious diseases such as pneumonia and diarrhea exhibit age-specific incidence patterns that peak in the post-neonatal period (Liu et al., 2016; Perin et al., 2022b; Sepanlou et al., 2022). Malaria mortality exhibits complex age patterns varying by endemicity level. Deaths in children under 3 months are rarely attributed to malaria, with peak mortality occurring at $1 - 4$ years in endemic areas (Abdullah et al., 2007). These patterns highlight why broad age categories can obscure critical epidemiological insights necessary for targeted interventions.

Despite the epidemiological need for age precision, researchers encounter mortality data with varying levels of age detail, ranging from precise age-at-death records measured in days to broad categories spanning multiple years. This heterogeneity creates a complex

decision matrix where researchers must choose between incorporating incomplete records with measurement error or excluding them and introducing selection bias that systematically underrepresents vulnerable populations. Recent evidence from WHO mortality databases reveals that less than 65% of deaths are recorded in many low-income countries, with age data being particularly problematic (World Health Organization, 2021b). This creates a scenario where populations with the highest mortality burdens face the greatest likelihood of exclusion from analyses.

The equity implications of age data handling decisions extend beyond technical statistical considerations. Research from sub-Saharan Africa demonstrates that household wealth, maternal education, and rural residence significantly predict both under-five mortality rates and data completeness (Chowdhury et al., 2017; Van Malderen et al., 2019; Fagbamigbe et al., 2022). Children from poorer households are systematically more likely to have incomplete age documentation at death, creating biases that compound existing health inequities (Memon et al., 2023). Studies examining Health and Demographic Surveillance Sites (HDSS) have found that excluding records with incomplete age data leads to a systematic underestimation of mortality rates in the most vulnerable populations (Erchick et al., 2023). This presents a methodological paradox where maintaining rigorous data quality standards may systematically exclude evidence from populations requiring intervention. The emerging data justice framework emphasizes balancing methodological rigor with representativeness, marking a departure from traditional approaches that prioritized internal validity over equity considerations (Shaw and Sekalala, 2023).

Current statistical approaches appear to have limitations in addressing this challenge. Although the most common approach, complete case analysis, shows inferior performance compared to principled methods, such as multiple imputation, when data are missing at random (Mukaka et al., 2016; Wells et al., 2013). However, the fraction of missing information emerges as a better predictor of efficiency gains than the raw proportion of missing data, with performance dependent on relationships between variables and

availability of auxiliary information (Madley-Dowd et al., 2019). Global health organizations have established standardized approaches while acknowledging persistent implementation challenges. The UN Inter-agency Group for Child Mortality Estimation (UN IGME) establishes minimum age categories distinguishing neonatal from post-neonatal deaths (Sharrow et al., 2022), with preference for six detailed categories (Mulick et al., 2022; Perin et al., 2022b). The WHO's 2022 Verbal Autopsy (VA) instrument maintains age-specific precision while simplifying data collection, and the ICD-11 implementation provides enhanced digital capabilities for handling age data (World Health Organization, 2017, 2022a; Fenton et al., 2024). The Institute for Health Metrics and Evaluation (IHME)'s Global Burden of Disease study employs six detailed age groups under five years, utilizing spatiotemporal Gaussian process regression with 7417 data sources (Global Burden of Disease Collaborative Network, 2020). Nevertheless, these standards acknowledge that nearly two-thirds of low- and middle-income countries lack reliable mortality data, with only approximately 60 countries maintaining fully functioning vital registration systems (Byass, 2009; Wyber et al., 2015).

The methodological landscape has evolved toward sophisticated Bayesian hierarchical models as powerful tools for handling incomplete age data. Recent innovations include flexible frameworks that impute suppressed age-specific death counts while accounting for multilevel data structures, zero counts, and right-skewed distributions (Arató et al., 2006). These models demonstrate accuracy within $2 - 4\%$ of national totals when validated against complete data (Kaul et al., 2024). For VA applications, frameworks like `InSilicoVA` use MCMC methods to assign COD while explicitly accounting for age uncertainty (Li et al., 2023; McCormick et al., 2016b). The `InterVA-5` system integrates age-dependent conditional probabilities for symptom-cause relationships, with validation studies from the PHMRC demonstrating robust performance across diverse settings (Byass et al., 2019; Zhu et al., 2025). Recent research has developed Bayesian age-category reconciliation frameworks that explicitly model probabilistic relationships between different age grouping schemes (Fan et al., 2023). Additional innovations include

harmonization approaches using cross-validation and simulation for robust parameter estimation (Guillot et al., 2022) and discrete hazards modeling that accommodates temporal heterogeneity in age reporting across different data collections (Burstein et al., 2018). These approaches underscore the need for flexible analytical tools that can accommodate varying data quality and reporting conventions across different surveillance systems and geographic contexts.

Despite these methodological advancements, a critical gap remains in addressing cause-specific mortality estimation when combining data sources with inconsistent age groupings (Gilbert et al., 2023; Amek, 2013; Hallett et al., 2010). When combining data from multiple sources, records may report different levels of age detail. Some sources provide precise age groups such as $1-11$ months and $12-59$ months, while others indicate broader categories such as $1-59$ months (Burstein et al., 2018; Gilbert et al., 2023; Schumacher et al., 2022). This challenge is particularly acute for cause-specific analysis because age distribution patterns vary substantially by COD. Systematic exclusion of records with incomplete age information may disproportionately affect data collected from resource-constrained settings, potentially introducing systematic bias into global health estimates and undermining efforts to monitor progress toward mortality reduction targets (Lawn et al., 2016; Black et al., 2016; You et al., 2015).

This chapter addresses this gap by developing and evaluating a framework for COD estimation with incomplete age classifications. We propose a Bayesian approach that integrates both fully observed and partially observed age data through a mixing parameter and evaluates its performance against traditional approaches that discard incomplete records. The method is assessed using VA data from multiple countries, with systematic masking of age information to simulate real-world scenarios of incomplete reporting.

5.2 Method

5.2.1 COD Estimation with Partial Age Observability

We develop a Bayesian framework to address incomplete age stratification in VA data by considering L distinct COD across $D+f$ administrative divisions, where D divisions provide only aggregated age information and f divisions supply complete age stratification. The covariate information is represented through matrices $\mathbf{X}^p \in \mathbb{R}^{D\times C}$ and $\mathbf{X}^f \in \mathbb{R}^{f\times C}$ for partially and fully observed divisions respectively, encompassing C demographic and socioeconomic predictors.

The observed death counts are structured as $\mathbf{Y}^p_{1-11}, \mathbf{Y}^p_{12-59}, \mathbf{Y}_{1-59} \in \mathbb{Z}^{D\times L}$ for divisions with incomplete age detail, and $\mathbf{Y}^f_{1-11}, \mathbf{Y}^f_{12-59} \in \mathbb{Z}^{f\times L}$ for divisions with complete age stratification. The selection of age groups reflects both epidemiological principles and practical considerations. The $1-59$ month aggregate corresponds to the under-five mortality indicator, which remains the most widely reported measure of pediatric mortality across health information systems (You et al., 2015). The subdivision at 12 months captures the critical epidemiological transition between late infant mortality ($1-11$ months) and early childhood mortality ($12-59$ months), periods characterized by distinct COD profiles and risk factor patterns (Lawn et al., 2014).

The proposed model introduces two core parameter sets that enable flexible modeling of age-specific mortality patterns. The mixing proportion parameter $\alpha_{1-11} \in (0,1)$ quantifies the relative contribution of early infant mortality ($1-11$ months) within the broader infant-toddler category ($1-59$ months). This parameter enables decomposition of aggregated age groups based on patterns observed in fully stratified data. Cause-specific regression coefficient matrices $\boldsymbol{\beta}_{1-11}, \boldsymbol{\beta}_{12-59} \in \mathbb{R}^{C\times L}$ capture the relationship between covariates and mortality probabilities for each age stratum. The mixing proportion α_{1-11} constitutes the central innovation of our modeling approach, enabling probabilistic inference about the underlying age composition of incomplete mortality data.

For administrative divisions with incomplete age stratification, the model employs a

hierarchical structure that links observed aggregated counts to underlying age-specific distributions:

$$\begin{aligned} \text{logit}(p^p_{1-11}[d]) &= \mathbf{X}^p[d] \cdot \beta_{1-11}, \\ \text{logit}(p^p_{12-59}[d]) &= \mathbf{X}^p[d] \cdot \beta_{12-59}, \\ p_{1-59}[d,l] &= \alpha_{1-11} \cdot p^p_{1-11}[d,l] + (1-\alpha_{1-11}) \cdot p^p_{12-59}[d,l]. \end{aligned} \tag{5.1}$$

Divisions providing complete age detail contribute direct information about age-specific mortality patterns through standard logistic regression formulations:

$$\begin{aligned} \text{logit}(p^f_{1-11}[s]) &= \mathbf{X}^f[s] \cdot \beta_{1-11}, \\ \text{logit}(p^f_{12-59}[s]) &= \mathbf{X}^f[s] \cdot \beta_{12-59}. \end{aligned}$$

The core contribution of this modeling framework lies in the weighted mixture formulation presented in equation (5.1), which transforms the problem of disaggregating incomplete age groups into a Bayesian inference task. This approach leverages complementary information available across different data sources, allowing age-specific COD patterns from fully stratified divisions to inform the decomposition of aggregated data from partially observed divisions.

To capture the discrete count nature of mortality data across COD categories, we employ multinomial likelihood functions:

$$\begin{aligned} \mathbf{Y}^p_a[d] &\sim \text{Multinomial}(p^p_a[d]), \\ \mathbf{Y}^f_a[s] &\sim \text{Multinomial}(p^f_a[s]), \\ \mathbf{Y}_b[d] &\sim \text{Multinomial}(p_b[d]), \end{aligned}$$

where $a \in \{1-11 \text{ months}, 12-59 \text{ months}\}$ and $b \in \{1-59 \text{ months}\}$.

Following established Bayesian modeling practices, we adopt minimally informative prior distributions that allow observed data to dominate posterior inference while providing necessary regularization:

$$\begin{aligned} \alpha_{1-11} &\sim \text{Beta}(1,1), \\ \beta_a[j] &\sim \mathcal{N}(0, 0.5^2) \text{ for } j \in \{1, \ldots, C\}, \end{aligned}$$

where $a \in \{1-11 \text{ months}, 12-59 \text{ months}\}$

The uniform Beta prior on the mixing proportion reflects the absence of strong prior assumptions about age distribution within masked categories, consistent with the principle of empirical evidence driving statistical inference (Jaynes, 2003). The normal priors on regression coefficients provide mild regularization while maintaining sufficient flexibility for data-driven parameter estimation, following recommendations for weakly informative priors in epidemiological modeling (Lemoine, 2019). Statistical inference is performed using the Stan probabilistic programming language (Carpenter et al., 2017).

5.3 Experiments

5.3.1 Verbal Autopsies in Demographic and Health Surveys

The empirical analysis utilizes VA data from the Demographic and Health Surveys (DHS) program, which provides standardized, nationally representative demographic and health information across low- and middle-income countries (LMICs) (Corsi et al., 2012). This analysis incorporates VA data from five distinct survey implementations: *Bangladesh* 2011, *Bangladesh* 2017–18, *Bangladesh* 2022, *Pakistan* 2006–07, and *Ghana* 2008. These surveys constitute the complete set of DHS implementations that incorporated VA modules for under-five mortality assessment during their respective data collection periods. The sub-national administrative disaggregation varies across countries according to their respective governmental structures. *Bangladesh* 2011 encompasses seven administrative divisions: *Barisal*, *Chittagong*, *Dhaka*, *Khulna*, *Rajshahi*, *Rangpur*, and *Sylhet*. *Bangladesh* 2017–18 expands this framework to eight divisions through the incorporation of *Mymensingh* division. *Bangladesh* 2022 maintains this eight-division administrative structure. *Pakistan* 2006–07 encompasses four provincial divisions: *Punjab*, *Sindh*, *NWFP* (subsequently renamed *Khyber Pakhtunkhwa*), and *Balochistan*. *Ghana* 2008 employs a simplified binary classification system distinguishing *Urban* and *Rural* geographical areas.

The data processing methodology involves systematic aggregation of sub-national mortality records into two primary age stratifications: 1–11 months and 12–59 months. This age categorization addresses a fundamental analytical challenge in which certain mortality records contain complete age-specific information. In contrast, others report only aggregate age ranges (e.g., 1–59 months), thereby creating a heterogeneous dataset characterized by both fully observed and partially observed age classifications. This methodological approach enables comprehensive utilization of available data while accounting for variations in data completeness across different administrative units and survey implementations.

The original COD classifications demonstrate substantial heterogeneity across countries and survey years, reflecting variations in local disease patterns, epidemiological contexts, and coding methodologies. *Bangladesh* 2011 employed nineteen distinct cause categories, encompassing conditions ranging from neonatal tetanus and congenital abnormalities to pneumonia, diarrhea, and various undetermined classifications. *Bangladesh* 2017–18 and 2022 utilized a more standardized eleven-category system that consolidated similar conditions and maintained consistent coding frameworks across these two recent survey periods. *Pakistan* 2006–07 implemented an extensive 128-category classification system, providing highly granular cause specification that included detailed neonatal conditions, infectious diseases, and specific injury mechanisms. *Ghana* 2008 employed cause categories similar to those utilized in the *Bangladesh* surveys, maintaining consistency with established DHS classification protocols.

The analysis standardizes these diverse original classifications into four broad cause categories through systematic mapping procedures designed to ensure analytical consistency and comparability across datasets. The "infectious diseases" category encompasses pneumonia, malaria, meningitis, and diarrhea-related mortality. The "neonatal and nutritional" category includes birth complications, preterm birth, neonatal sepsis, congenital anomalies, and malnutrition-related conditions. The "injuries" category encompasses accidents, drowning, violence, and other external causes of death. The "other" category

includes unclassified, unknown, or indeterminate causes that cannot be reliably assigned to the three primary categories. This standardization framework facilitates cross-country comparative analysis while preserving the clinical and epidemiological relevance of the underlying cause classifications.

Three testing scenarios are designed to assess model performance under varying data availability conditions. Case 1 utilizes complete data where all administrative units provide fully stratified age information across both age groups, serving as the reference standard. Case 2 incorporates mixed data availability, combining 80% fully observed records with 20% partially observed data where age stratification is masked. Case 3 implements complete-case analysis, utilizing only the 80% of fully observed data while discarding records with missing age specifications. This synthetic data generation process employed randomized masking procedures to create controlled experimental conditions while preserving the underlying statistical properties of the original datasets. We randomly selected 20% of complete records. We synthetically masked their age stratification, systematically varying the true mixing proportion parameter α_{1-11} across the range 0.1 to 0.9 to evaluate model robustness across diverse age distribution scenarios. This approach enables precise assessment of parameter recovery accuracy while maintaining the authentic correlation structures and covariate relationships present in the empirical data (Glynn et al., 1993).

Model performance assessment employed a comprehensive evaluation framework spanning multiple dimensions of statistical accuracy and distributional fidelity. Point estimate precision was quantified through Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) (Hastie et al., 2009). Distributional accuracy was assessed using information-theoretic measures, including Kullback-Leibler (KL) Divergence and Jensen-Shannon Divergence, which capture the degree of similarity between estimated and true probability distributions (Klir and Wierman, 1999). We evaluated correlation preservation through Mean Correlation coefficients and parameter estimation accuracy through direct comparison of recovered mixing proportions against known true values, ensuring a

comprehensive assessment of both predictive performance and inferential validity. This approach aligns with established practices for evaluating statistical models in epidemiological contexts where both predictive accuracy and uncertainty quantification are critical for policy applications (Kneib, 2015).

5.3.2 Empirical Results

Performance Across Data Completeness Scenarios

The comparison of modeling approaches under varying data availability conditions demonstrates the advantages of incorporating partially observed data rather than complete-case analysis. Figure 5.1 presents the proportion of evaluation metrics where the proposed approach (Case 2) achieves better performance relative to the data exclusion strategy (Case 3) across different age distribution scenarios. The results show that leveraging incomplete data through our modeling approach outperforms the conventional practice of discarding records with missing information.

The performance differences between modeling approaches exhibit patterns that reflect underlying epidemiological characteristics of under-five child mortality. Across most age distribution scenarios, the proposed framework achieves better performance on more than half of the evaluation metrics, with performance gains being particularly notable for the $12-59$ month age stratum. This pattern likely reflects the greater statistical stability and larger sample sizes typically observed in the older infant and toddler mortality categories, enabling more robust parameter estimation through the Bayesian approach (Wakefield et al., 2013). The proposed approach achieves consistent performance improvements across all evaluation metrics and age distribution scenarios, suggesting that the benefits of incorporating incomplete data are particularly evident when analyzing mortality patterns in older infant and toddler populations.

The detailed assessment of performance improvements across individual evaluation metrics reveals varied but consistently positive effects of the proposed approach. Fig-

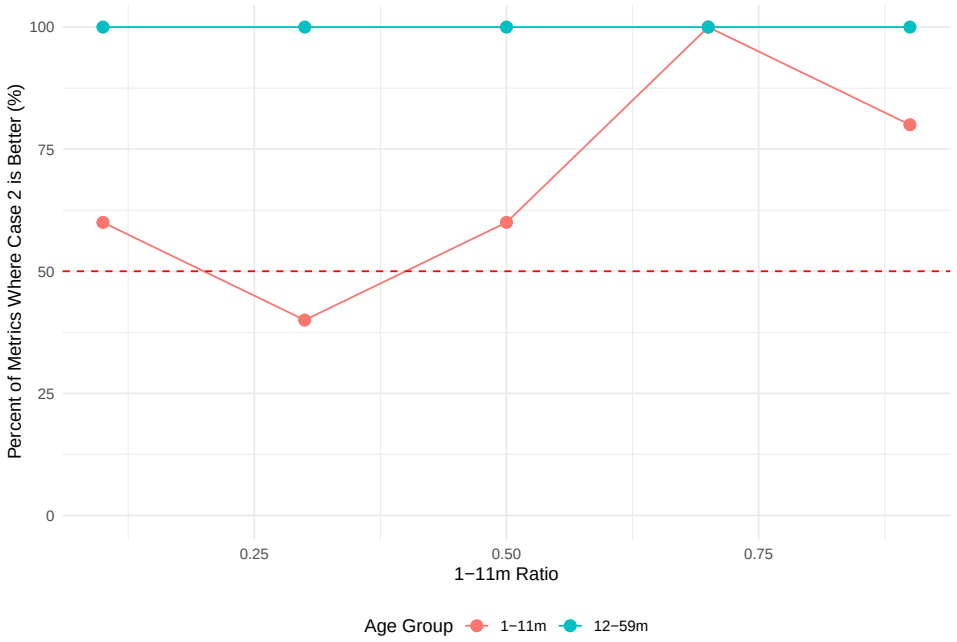


Figure 5.1: Performance comparison between Case 2 and Case 3 across age groups and ratio conditions. The analysis presents the percentage of evaluation metrics where Case 2 demonstrates superior performance relative to Case 3, with values above 50% indicating general superiority of Case 2. Results show contrasting patterns between age groups: the $12-59$ month age group consistently favors Case 2 across all ratio conditions while the $1-11$ month age group exhibits variable performance. The dashed horizontal line at 50% provides a reference threshold for determining overall method superiority.

ure 5.2 examines performance differences between the two approaches using five metrics: Jensen-Shannon Divergence, KL Divergence, MAE, Mean Correlation, and RMSE. Results demonstrate substantial variation in relative performance depending on both the evaluation metric and the age group under consideration. The $12 - 59$ month age group consistently shows positive improvements for Case 2 across most metrics and ratio conditions, with particularly strong performance in Jensen-Shannon Divergence and Mean Correlation measures. The $1 - 11$ month age group exhibits more variable performance patterns, with Case 2 showing significant deterioration at the 0.5 ratio condition across multiple metrics, particularly Mean Correlation and RMSE, where performance declines exceed 400%. These findings suggest that the optimal methodology selection depends critically on both the target age population and the specific analytical requirements of the application (Carpenter et al., 2023). Information-theoretic measures, particularly the KL divergence, exhibit significant performance gains under the proposed framework. This finding is important because KL divergence quantifies the information loss associated with using an approximate distribution instead of the true distribution, suggesting that the proposed approach substantially reduces information loss compared to complete-case analysis (Cover, 1999).

Regression Parameter Estimation

Figure 5.3 demonstrates that the age-specific regression coefficients β_{1-11} and β_{12-59} are estimated with greater precision under the proposed approach compared to complete-case analysis across most age distribution scenarios. This improvement in parameter estimation quality directly translates into enhanced inferential capacity for epidemiological research applications where accurate quantification of covariate effects is essential.

The proximity of proposed model parameter estimates to the reference standard established using complete data provides validation of the proposed approach. This convergence toward the true parameter values demonstrates that the proposed modeling framework effectively recovers the underlying epidemiological associations that would be

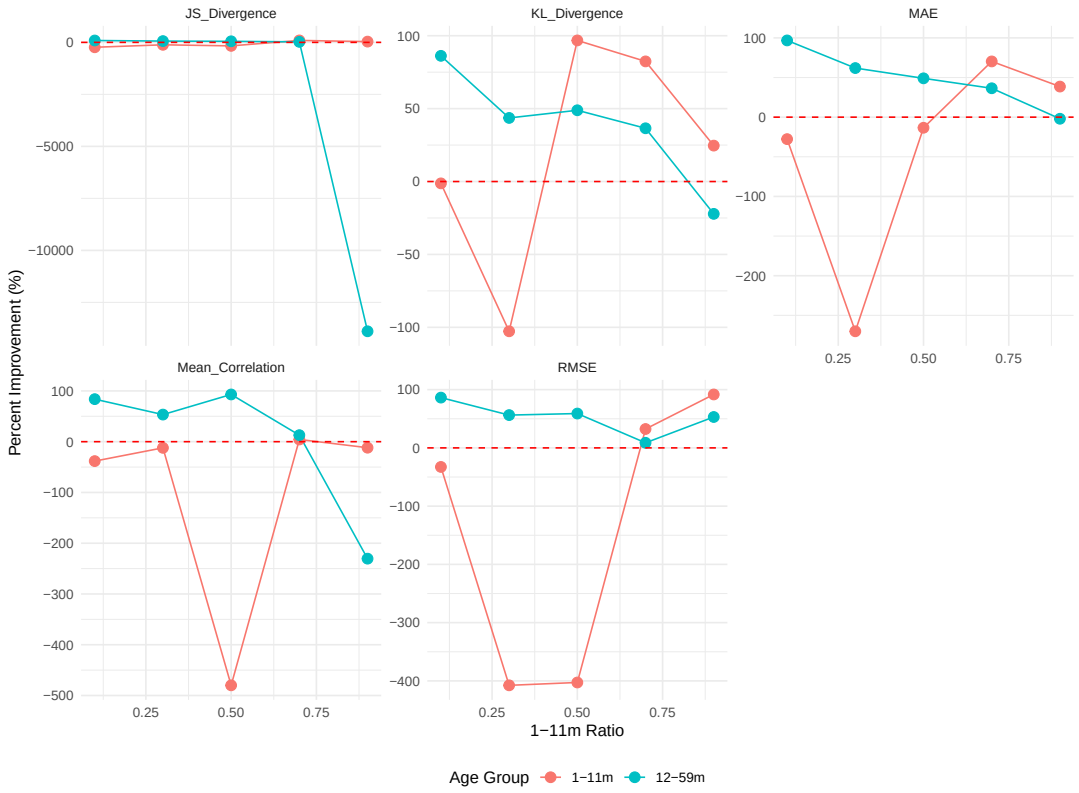


Figure 5.2: Percentage improvement of Case 2 methodology relative to Case 3 across multiple evaluation metrics and ratio conditions. Positive values indicate superior performance by Case 2, while negative values favor Case 3.

observable under ideal data conditions. At the same time, complete-case analysis systematically deviates from these true relationships due to information loss.

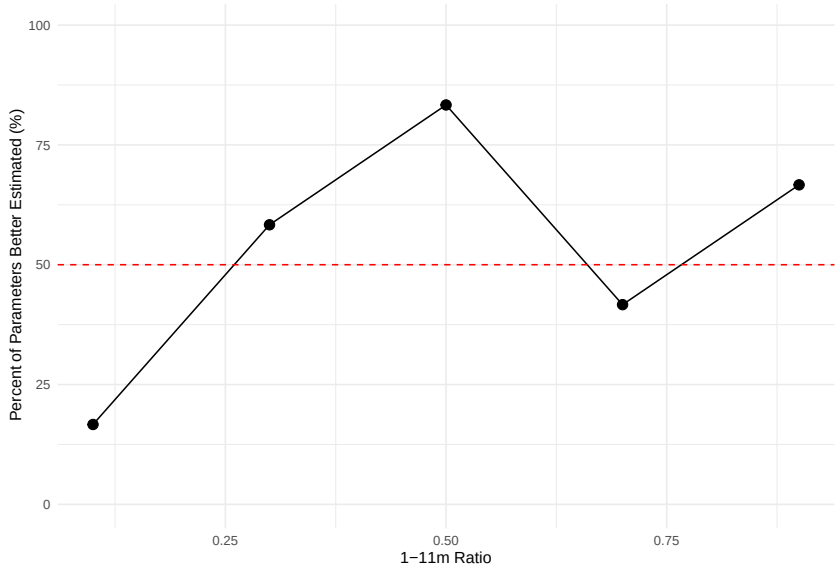
The quantitative assessment of parameter estimation error supports these findings through direct measurement of absolute deviations from reference values. Figure 5.4 illustrates that the proposed approach consistently yields lower parameter estimation errors across all evaluated scenarios, providing further empirical evidence for the enhanced inferential properties of the proposed model.

Mixing Parameter Estimation

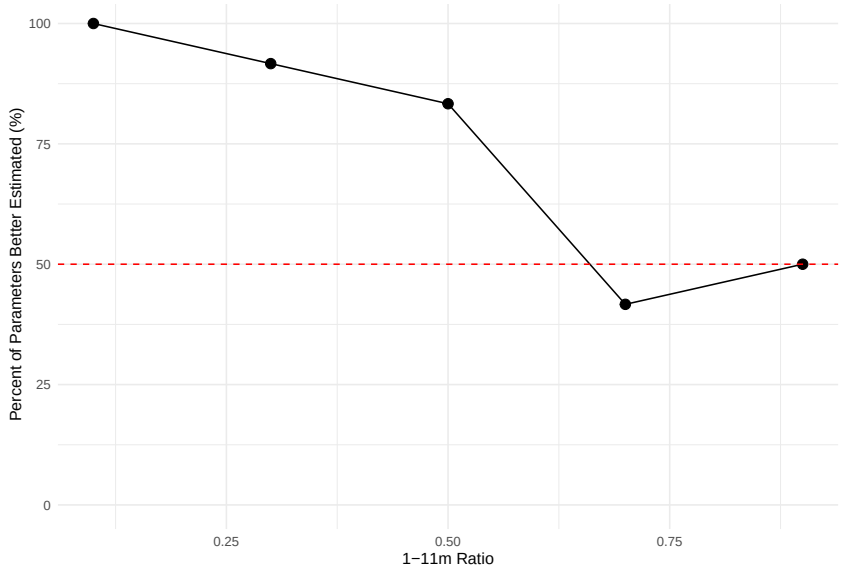
Figure 5.5 assesses the precision with which the model recovers the true age distribution parameter α_{1-11} from observed data patterns, providing direct evidence for the model's capacity to learn latent age structures from heterogeneous data sources. This capability is fundamental to the practical utility of the approach in real-world applications where the true age distribution within masked categories is unknown (Richardson and Green, 1997). The empirical results demonstrate satisfactory parameter recovery performance, with estimated mixing proportions clustering closely around the ideal diagonal relationship between true and estimated values. This pattern indicates that the proposed Bayesian inference framework successfully learns the underlying age distribution structure from the observed data patterns, enabling reliable disaggregation of masked categories even when the true age composition is unknown a priori.

Cause-Specific Prediction Accuracy

Figures 5.6 and 5.7 present comprehensive comparisons of predicted versus empirical COD probabilities. The varied performance patterns observed across different COD reflect the distinct epidemiological characteristics and statistical properties of varying mortality patterns. The improvement in prediction accuracy achieved through the proposed approach is evident in the tighter clustering of estimated probabilities around the diagonal reference line, indicating reduced bias and enhanced precision across multiple



(a) Percentage of β_{1-11} better estimated in Case 2 across different ratios.



(b) Percentage of β_{12-59} better estimated in Case 2 across different ratios.

Figure 5.3: Parameter estimation accuracy comparison between Case 2 and Case 3 across different ratio conditions. Panel (a) shows the percentage of β_{1-11} parameters where Case 2 provides better estimates. Panel (b) presents results for β_{12-59} parameters. The horizontal reference line at 50% indicates the threshold for overall method superiority.

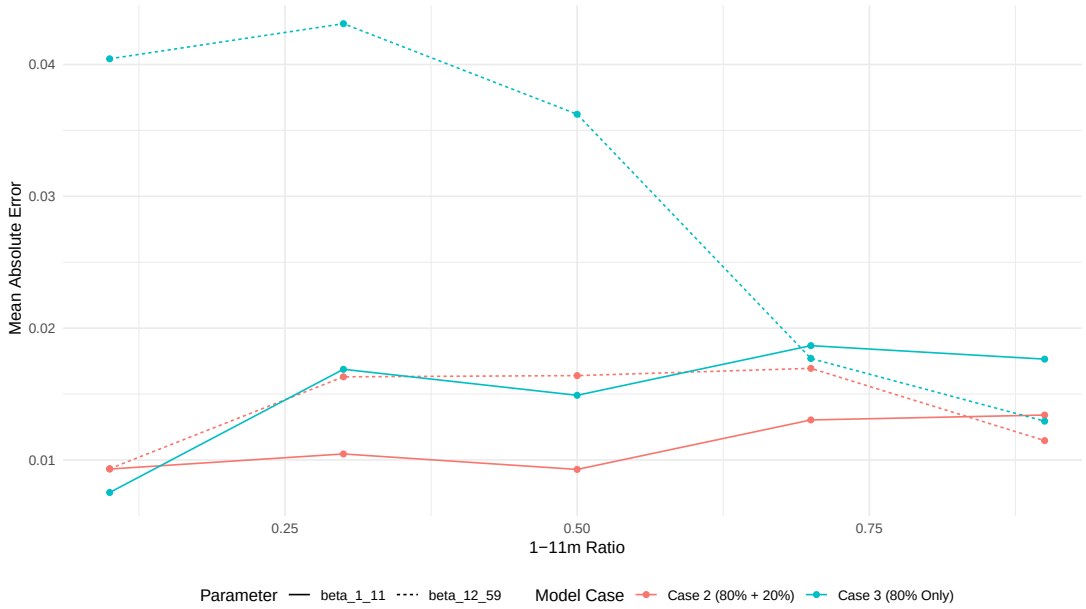


Figure 5.4: Mean absolute parameter error comparison between Case 2 and Case 3 across ratio conditions. Case 2 (red lines) consistently demonstrates lower estimation errors than Case 3 (blue lines) for β_{1-11} parameters across all ratios. For β_{12-59} parameters, Case 2 maintains superior performance at lower ratios, with the two approaches converging at higher ratio values.

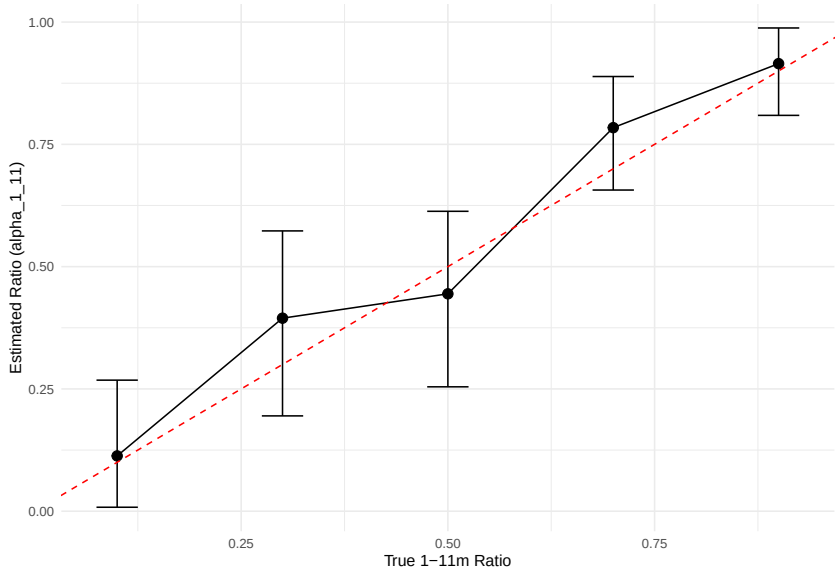


Figure 5.5: Estimation of α_{1-11} with error bars across ratio conditions. Estimated values are plotted against true values, with the diagonal line indicating perfect estimation.

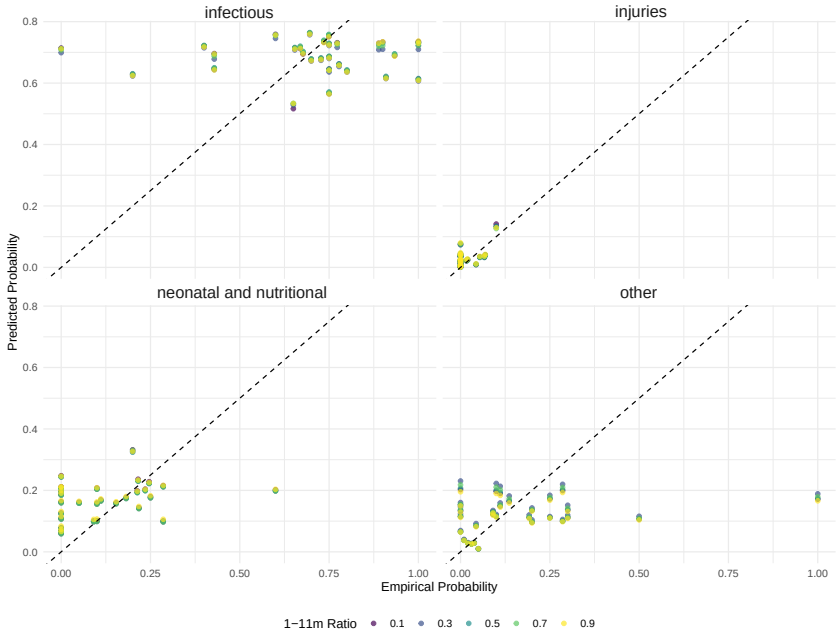
CODs. The performance improvements observed for infectious diseases and injuries in the $12-59$ month age group merit particular attention given their prominence in global child mortality patterns. These cause categories represent major contributors to preventable childhood deaths, and the enhanced prediction accuracy achieved through the proposed modeling directly translates into improved capacity for epidemiological surveillance and intervention planning in these critical areas (Liu et al., 2016).

5.4 Discussions

The empirical validation presented in this chapter provides evidence for the advantages of Bayesian approaches in mortality surveillance applications with heterogeneous data completeness. The improvements observed across multiple evaluation metrics show that the proposed framework effectively leverages information contained within partially observed age information, thereby reducing the information loss associated with conventional complete-case analysis strategies. This finding aligns with established principles in missing data methodology while demonstrating specific advantages for epidemiological applications where data exclusion can compromise statistical power and population representativeness (White et al., 2011). It suggests that the benefits of incorporating incomplete data extend beyond gains in sample size to improvements in the quality of epidemiological inference (Molenberghs et al., 2014).

The observed variation in performance improvements across different age distribution scenarios reveals essential characteristics of the proposed methodology and guides practical implementation. Under conditions of extreme age distribution skewness within masked categories, the proposed approach exhibits some reduction in estimation precision while maintaining better performance relative to data exclusion strategies. The capacity for data-driven learning of latent age structure represents a methodological advantage that distinguishes the proposed approach from conventional imputation strategies that rely on external assumptions or auxiliary data sources. The adaptive characteristics of the Bayesian framework make it suitable for applications in diverse global health contexts

(a) Predicted vs. Empirical probabilities for Case 2 (1-11m).



(b) Predicted vs. Empirical probabilities for Case 3 (1-11m).

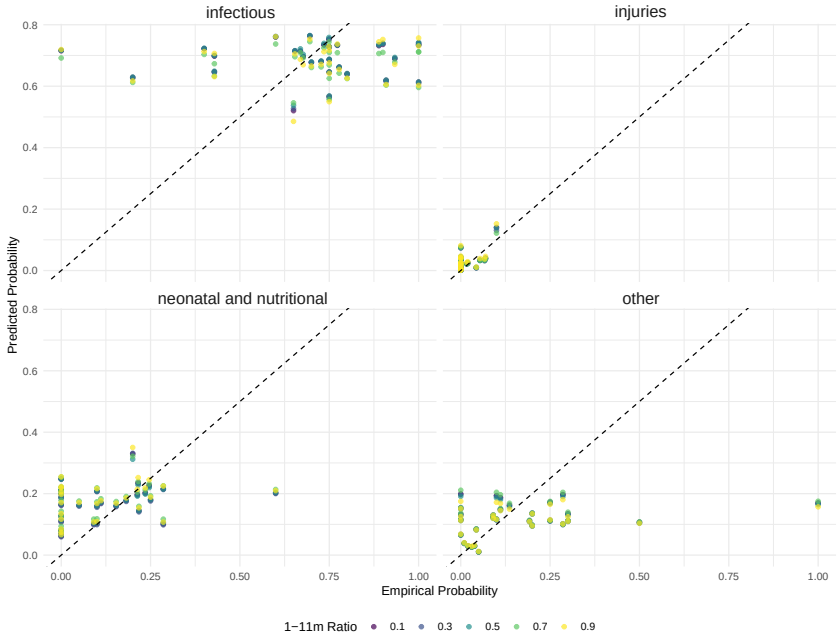
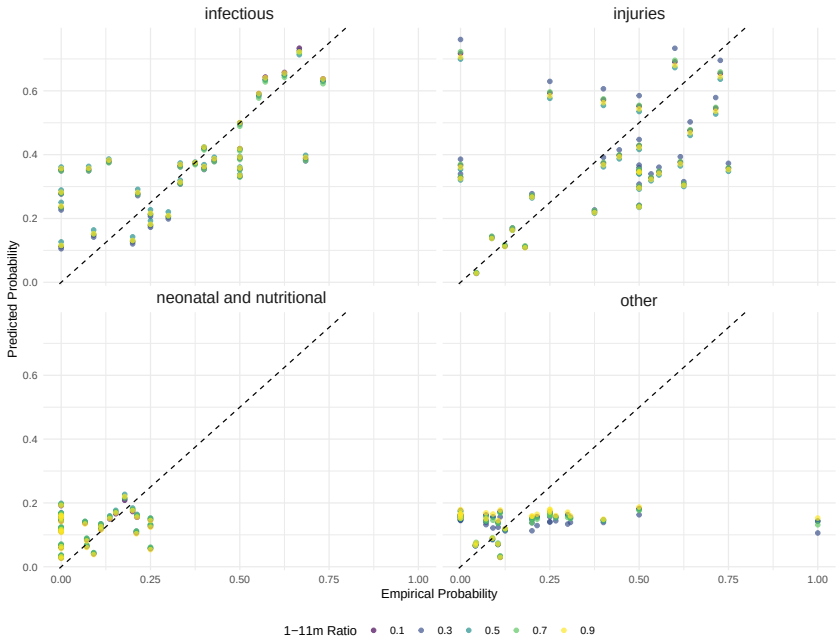


Figure 5.6: Comparison of predicted and empirical probabilities between Case 2 and Case 3 for the $1-11$ month age group, with the diagonal line indicating perfect calibration.

(a) Predicted vs. Empirical probabilities for Case 2 (12-59m).



(b) Predicted vs. Empirical probabilities for Case 3 (12-59m).

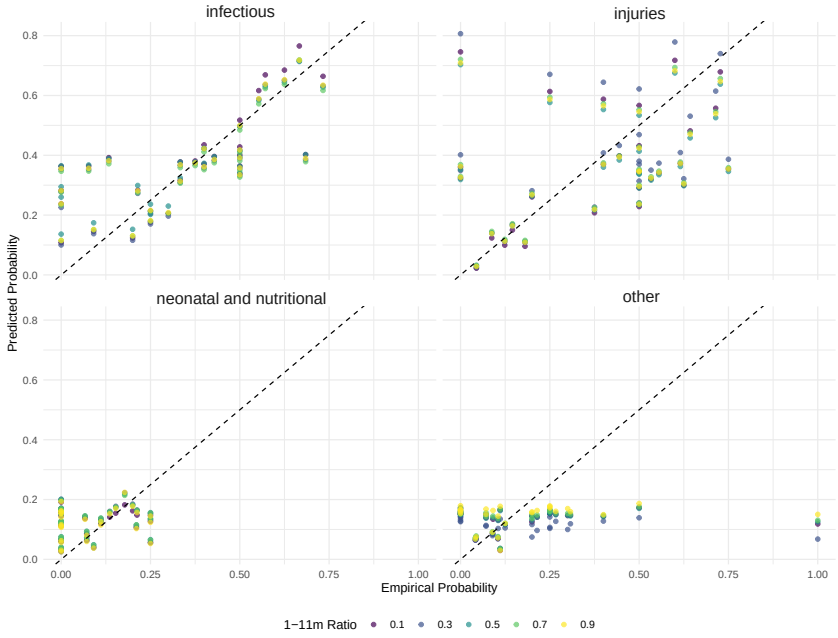


Figure 5.7: Comparison of predicted and empirical probabilities between Case 2 and Case 3 for the $12-59$ month age group, with the diagonal line indicating perfect calibration.

where demographic patterns are variable across geographic regions, temporal periods, or socioeconomic strata. This flexibility is particularly relevant for mortality surveillance in LMICs, where data systems may exhibit heterogeneous reporting practices and where external demographic information for imputation purposes may be limited or unreliable (Hill et al., 2012).

Several important limitations must be acknowledged. The current validation is based on a simplified scenario with only two age groups and assumes that data is missing at random. Real-world applications may involve more complex age stratifications and systematic patterns of missing data, which could impact model performance. The method also requires sufficient complete data to learn the underlying age distribution patterns, which may not always be available in practice. Additionally, the proposed model assumes that the relationship between covariates and mortality patterns is consistent across regions, regardless of whether data are complete or incomplete. This assumption may be violated when data completeness is systematically related to health system capacity or other factors that also influence mortality patterns.

Building on this proof-of-concept study, several important extensions warrant investigation. The framework could be expanded to handle more complex age stratifications beyond the two-group scenario tested here. Investigation of performance under different missing data patterns would help understand when the method works best and when it may struggle. The proposed approach could be extended to incorporate multiple demographic variables simultaneously, such as handling missing information about both age and location. Testing the method with larger datasets would help evaluate computational requirements. Future work should also examine how the method performs when the assumption of consistent covariate effects is violated and develop diagnostic tools to identify when this assumption may be problematic. Finally, applying this approach to real-world child mortality data from various countries would provide valuable insights into practical implementation challenges and benefits.

Chapter 6

CONCLUDING REMARKS

This dissertation has developed and explored four interconnected methodological frameworks that address fundamental challenges in VA data analysis. These contributions improve how researchers collect, integrate, and analyze COD data in settings where medical certification is unavailable.

Several limitations provide opportunities for future research development. Current approaches rely on relatively simple analytical models compared to state-of-the-art methods available in the literature. Integration with more sophisticated algorithms represents a key area for advancement that could enhance overall performance across all frameworks. The methodologies do not fully utilize the extensive domain knowledge embedded in traditional verbal autopsy design. Logical relationships and dependencies among questions could inform more effective implementations, particularly for adaptive questioning strategies that reflect clinical reasoning patterns.

A critical finding across this research concerns the complex relationship between predictive accuracy and inferential validity. Superior classification performance does not necessarily translate to enhanced parameter estimation quality. This observation has significant implications for model selection in applied epidemiological contexts and underscores the need for rigorous evaluation criteria that extend beyond traditional accuracy metrics.

Future research should prioritize integrated frameworks that address multiple sources of uncertainty simultaneously. Joint modeling approaches could account for both age and cause-of-death misclassifications, recognizing interdependencies between different types of classification errors to yield more accurate mortality estimates. Ensemble modeling

strategies would enhance robustness against model misspecification while providing more reliable uncertainty quantification across all methodological components. Population-specific adaptation represents another important research direction. The frameworks require extension to account for systematic variation in symptom-cause relationships across different healthcare contexts and demographic populations. Hierarchical modeling approaches or transfer learning methodologies can effectively address these variations. Human factors research constitutes a vital priority with direct implications for data quality and respondent welfare. Systematic investigation should examine interviewer and respondent responses to adaptive questioning, quantify the differential emotional burden across question types, and assess population-specific bias patterns. Comprehensive cost-effectiveness models that consider question-specific costs, respondent burden, and accuracy trade-offs would provide practical guidance for deployment decisions in resource-limited settings.

The frameworks developed in this dissertation strengthen health information systems in regions where the child mortality burden remains highest, enabling more efficient data integration, reducing collection burden, and ensuring valid inference with emerging computational approaches. This supports evidence-based public health decision-making in resource-constrained environments. These statistical frameworks extend beyond VA applications to broader survey-based inference problems. The approaches for handling incomplete data, optimizing sequential data collection, and ensuring valid inference with predicted outcomes apply across multiple research domains. This research also highlights a critical distinction between predictive accuracy and inferential validity, with significant implications for the growing integration of machine learning and artificial intelligence approaches in scientific applications. As automated predictors become increasingly prevalent in research contexts, these frameworks guide maintaining statistical rigor while leveraging computational advances.

BIBLIOGRAPHY

Abdullah, S., Adazu, K., Masanja, H., Diallo, D., Hodgson, A., Ilboudo-Sanogo, E., Nhacolo, A., Owusu-Agyei, S., Thompson, R., Smith, T., et al. (2007). Patterns of age-specific mortality in children in endemic areas of sub-saharan africa. *Defining and Defeating the Intolerable Burden of Malaria III: Progress and Perspectives: Supplement to Volume 77 (6) of American Journal of Tropical Medicine and Hygiene*.

AbouZahr, C., De Savigny, D., Mikkelsen, L., Setel, P. W., Lozano, R., Nichols, E., Notzon, F., and Lopez, A. D. (2015). Civil registration and vital statistics: progress in the data revolution for counting and accountability. *The Lancet*, 386(10001):1373–1385.

Adeloye, D., Bowman, K., Chan, K. Y., Patel, S., Campbell, H., and Rudan, I. (2018). Global and regional child deaths due to injuries: an assessment of the evidence. *Journal of global health*, 8(2):021104.

Agresti, A. (2013). *Categorical data analysis*. John Wiley & Sons.

Ahn, K. W. and Chan, K.-S. (2010). Efficient markov chain monte carlo with incomplete multinomial data. *Statistics and computing*, 20(4):447–456.

Ahn, K. W., Chan, K.-S., Bai, Y., and Kosoy, M. (2007). A note on bayesian inference with incomplete multinomial data with applications for assessing the spatio-temporal variation in pathogen-variant diversity. Technical report, Technical report 381. http://www. stat. uiowa. edu/techrep/tr381. pdf, The ....

Ahn, K. W., Chan, K.-S., Bai, Y., and Kosoy, M. (2010). Bayesian inference with incomplete multinomial data: A problem in pathogen diversity. *Journal of the American Statistical Association*, 105(490):600–611.

Alam, K. and Mitra, A. (1986). An empirical bayes estimate of multinomial probabilities. *Communications in Statistics-Theory and Methods*, 15(10):3103–3127.

Albert, J. H. (1985). Bayesian estimation methods for incomplete two-way contingency tables using prior belief of association. in "bayesian statistics"(jm bernardo,... and afm smith, eds.). *Amsterdam, North Holland*, 2:589–602.

Albert, J. H. (1987). Empirical bayes estimation in contingency tables. *Communications in Statistics-Theory and Methods*, 16(8):2459–2485.

Albert, J. H. and Gupta, A. K. (1983). Bayesian estimation methods for $2\times$ 2 contingency tables using mixtures of dirichlet distributions. *Journal of the American Statistical Association*, 78(383):708–717.

Amek, N. O. (2013). *Bayesian spatio-temporal modelling of the relationship between mortality and malaria transmission in rural western Kenya*. PhD thesis, University_of_Basel.

Angelopoulos, A. N., Bates, S., Fannjiang, C., Jordan, M. I., and Zrnic, T. (2023a). Prediction-powered inference. *Science*, 382(6671):669–674.

Angelopoulos, A. N., Duchi, J. C., and Zrnic, T. (2023b). Ppi++: Efficient prediction-powered inference. *arXiv preprint arXiv:2311.01453*.

Antelman, G. R. (1972). Interrelated bernoulli processes. *Journal of the American Statistical Association*, 67(340):831–841.

Arató, N. M., Dryden, I. L., and Taylor, C. C. (2006). Hierarchical bayesian modelling of spatial age-dependent mortality. *Computational Statistics & Data Analysis*, 51(2):1347–1363.

Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2003). *Hierarchical modeling and analysis for spatial data*. Chapman and Hall/CRC.

Baqui, A. H., Darmstadt, G. L., Williams, E., Kumar, V., Kiran, T., Panwar, D., Srivastava, V., Ahuja, R., Black, R., and Santosham, M. (2006). Rates, timing and causes of neonatal deaths in rural india: implications for neonatal health programmes. *Bulletin of the World Health Organization*, 84:706–713.

Bareinboim, E. and Pearl, J. (2016). Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences*, 113(27):7345–7352.

Barocas, S. and Selbst, A. D. (2016). Big data's disparate impact. *Calif. L. Rev.*, 104:671.

Basu, D. and Pereira, C. A. d. B. (1982). On the bayesian analysis of categorical data: the problem of nonresponse. *Journal of Statistical Planning and Inference*, 6(4):345–362.

Berger, J. O. and Berger, J. O. (1985). *Bayesian analysis*. Springer.

Bernardo, J. M. and Smith, A. F. (2009). *Bayesian theory*, volume 405. John Wiley & Sons.

Besag, J., York, J., and Mollié, A. (1991). Bayesian image restoration, with two applications in spatial statistics. *Annals of the institute of statistical mathematics*, 43:1–20.

Biswas, S. S. (2023). Role of chat gpt in public health. *Annals of biomedical engineering*, 51(5):868–869.

Black, R. E., Levin, C., Walker, N., Chou, D., Liu, L., and Temmerman, M. (2016). Reproductive, maternal, newborn, and child health: key messages from disease control priorities 3rd edition. *The Lancet*, 388(10061):2811–2824.

Blanco, A., Perez, A., Casillas, A., and Cobos, D. (2020). Extracting cause of death from verbal autopsy with deep learning interpretable methods. *IEEE Journal of Biomedical and Health Informatics*.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877.

Blumenthal, S. (1968). Multinomial sampling with partially categorized data. *Journal of the American Statistical Association*, 63(322):542–551.

Botsis, T., Hartvigsen, G., Chen, F., and Weng, C. (2010). Secondary use of ehr: data quality issues and informatics opportunities. *Summit on translational bioinformatics*, 2010:1.

Bradley, R. A. and Terry, M. E. (1952). Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345.

Burstein, R., Wang, H., Reiner Jr, R. C., and Hay, S. I. (2018). Development and validation of a new method for indirect estimation of neonatal, infant, and child mortality trends using summary birth histories. *PLoS Medicine*, 15(10):e1002687.

Byass, P. (2009). The unequal world of health data. *PLoS medicine*, 6(11):e1000155.

Byass, P., Chandramohan, D., Clark, S. J., D'ambruoso, L., Fottrell, E., Graham, W. J., Herbst, A. J., Hodgson, A., Hounton, S., Kahn, K., et al. (2012). Strengthening standardised interpretation of verbal autopsy data: the new interva-4 tool. *Global health action*, 5(1):19281.

Byass, P., Hussain-Alkhateeb, L., D'Ambruoso, L., Clark, S., Davies, J., Fottrell, E., Bird, J., Kabudula, C., Tollman, S., Kahn, K., Linus, S., and Petzold, M. (2019). An integrated approach to processing WHO-2016 verbal autopsy data: the InterVA-5 model. *BMC Medicine*, 17(1):1–12.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of statistical software*, 76:1–32.

Carpenter, J. R., Bartlett, J. W., Morris, T. P., Wood, A. M., Quartagno, M., and Kenward, M. G. (2023). *Multiple imputation and its application*. John Wiley & Sons.

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika*, pages 465–480.

Cejudo, A., Casillas, A., Pérez, A., Oronoz, M., and Cobos, D. (2023). Cause of death estimation from verbal autopsies: Is the open response redundant or synergistic? *Artificial intelligence in medicine*, 143:102622.

Chandramohan, D., Fottrell, E., Leitao, J., Nichols, E., Clark, S. J., Alsokhn, C., Munoz, D. C., AbouZahr, C., Pasquale, A. D., Mswia, R., Choi, E., Baiden, F., Thomas, J., Lyatuu, I., Li, Z. R., Larbi-Debrah, P., Chu, Y., Cheburet, S., Sankoh, O., Badr, A. M., Fat, D. M., Setel, P., Jakob, R., and de Savigny, D. (2021). Estimating causes of death where there is no medical certification: evolution and state of the art of verbal autopsy. *Global Health Action*, 14(sup1):1982486.

Chang, Y.-P., Chiu, C.-Y., and Tsai, R.-C. (2019). Nonparametric CAT for CD in educational settings with small samples. *Applied Psychological Measurement*, 43(7):543–561.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357.

Chelsea Zhang, Sean J. Taylor, Curtiss Cobb, and Jasjeet Sekhon (2020). Active matrix factorization for surveys. *The Annals of Applied Statistics*, 14(3):1182–1206.

Chen, I. Y., Pierson, E., Rose, S., Joshi, S., Ferryman, K., and Ghassemi, M. (2021). Ethical machine learning in healthcare. *Annual review of biomedical data science*, 4(1):123–144.

Chen, P., Xin, T., Wang, C., and Chang, H.-H. (2012). Online calibration methods for the DINA model with independent attributes in CD-CAT. *Psychometrika*, 77(2):201–222.

Chen, T. and Fienberg, S. E. (1974). Two-dimensional contingency tables with both completely and partially cross-classified data. *Biometrics*, pages 629–642.

Chen, T. and Fienberg, S. E. (1976). The analysis of contingency tables with incompletely classified data. *Biometrics*, pages 133–144.

Cheng, Y. (2009). When cognitive diagnosis meets computerized adaptive testing: CD-CAT. *Psychometrika*, 74(4):619.

Cheng, Y. (2010). Improving Cognitive Diagnostic Computerized Adaptive Testing by Balancing Attribute Coverage: The Modified Maximum Global Discrimination Index Method. *Educational and Psychological Measurement*, 70(6):902–913.

Chipman, H. A., George, E. I., and McCulloch, R. E. (2010). Bart: Bayesian additive regression trees. *The Annals of Applied Statistics*, pages 266–298.

Choi, S. W., Grady, M. W., and Dodd, B. G. (2010). A New Stopping Rule for Computerized Adaptive Testing. *Educational and Psychological Measurement*, 70(6):1–17.

Choi, Y. and McClenen, C. (2020). Development of Adaptive Formative Assessment System Using Computerized Adaptive Testing and Dynamic Bayesian Networks. *Applied Sciences*, 10(22).

Chowdhury, A. H., Hanifi, S. M. A., Mia, M. N., and Bhuiya, A. (2017). Socioeconomic inequalities in under-five mortality in rural bangladesh: evidence from seven national surveys spreading over 20 years. *International journal for equity in health*, 16:1–7.

Clark, S., Wakefield, J., McCormick, T., and Ross, M. (2018a). Hyak mortality monitoring system: innovative sampling and estimation methods–proof of concept by simulation. *Global health, epidemiology and genomics*, 3:e3.

Clark, S. J., Li, Z., and McCormick, T. H. (2018b). Quantifying the contributions of training data and algorithm logic to the performance of automated cause-assignment algorithms for verbal autopsy. *arXiv preprint arXiv:1803.07141*.

Cohn, D. A., Ghahramani, Z., and Jordan, M. I. (1996). Active learning with statistical models. *Journal of Artificial Intelligence Research*, 4:129–145.

Congdon, P. D. (2010). *Applied Bayesian hierarchical methods*. Chapman and Hall/CRC.

Corsi, D. J., Neuman, M., Finlay, J. E., and Subramanian, S. (2012). Demographic and health surveys: a profile. *International journal of epidemiology*, 41(6):1602–1613.

Cover, T. M. (1999). *Elements of information theory*. John Wiley & Sons.

Cressie, N., Davis, A. S., Folks, J. L., and Folks, J. L. (1981). The moment-generating function and negative integer moments. *The American Statistician*, 35(3):148–150.

Cressie, N. and Read, T. R. (1984). Multinomial goodness-of-fit tests. *Journal of the Royal Statistical Society: Series B (Methodological)*, 46(3):440–464.

D. Wu (2019). Pool-Based Sequential Active Learning for Regression. *IEEE Transactions on Neural Networks and Learning Systems*, 30(5):1348–1359.

D'Ambruoso, L., Boerma, T., Byass, P., Fottrell, E., Herbst, K., Källander, K., and Mullan, Z. (2017). The case for verbal autopsy in health systems strengthening. *The Lancet Global Health*, 5(1):e20–e21.

Daniel Mimoso Paulino, C. and Alberto De Bragança Pereira, C. (1992). Bayesian analysis of categorical data informatively censored. *Communications in Statistics-Theory and Methods*, 21(9):2689–2705.

Danso, S., Atwell, E., and Johnson, O. (2013). Linguistic and statistically derived features for cause of death prediction from verbal autopsy text. In *Language Processing and Knowledge in the Web: 25th International Conference, GSCL 2013, Darmstadt, Germany, September 25-27, 2013. Proceedings*, pages 47–60. Springer.

Datta, A., Fiksel, J., Amouzou, A., and Zeger, S. L. (2021). Regularized Bayesian transfer learning for population-level etiological distributions. *Biostatistics*, 22(4):836–857.

Dempster, A. and Shafer, G. (1976). *A Mathematical Theory of Evidence*. Limited paperback editions. Princeton University Press.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Diaz, T., Strong, K. L., Cao, B., Guthold, R., Moran, A. C., Moller, A.-B., Requejo, J., Sadana, R., Thiyagarajan, J. A., Adebayo, E., et al. (2021). A call for standardised age-disaggregated health data. *The Lancet Healthy Longevity*, 2(7):e436–e443.

Dickey, J. M., Jiang, J.-M., and Kadane, J. B. (1987). Bayesian methods for censored categorical data. *Journal of the American Statistical Association*, 82(399):773–781.

Diggle, P. J., Moraga, P., Rowlingson, B., and Taylor, B. M. (2013). Spatial and spatio-temporal log-gaussian cox processes: Extending the geostatistical paradigm. *Statistical Science*, pages 542–563.

Dong, F. and Yin, G. (2018). Maximum likelihood estimation for incomplete multinomial data via the weaver algorithm. *Statistics and Computing*, 28:1095–1117.

Dor, L. E., Halfon, A., Gera, A., Shnarch, E., Dankin, L., Choshen, L., Danilevsky, M., Aharonov, R., Katz, Y., and Slonim, N. (2020). Active learning for BERT: An empirical study. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7949–7962.

D'Orazio, M., Di Zio, M., and Scanu, M. (2006). *Statistical matching: Theory and practice*. John Wiley & Sons.

Doshi, F., Miller, K., Van Gael, J., and Teh, Y. W. (2009). Variational inference for the indian buffet process. In *Artificial Intelligence and Statistics*, pages 137–144. PMLR.

Dunson, D. B. and Xing, C. (2009). Nonparametric Bayes Modeling of Multivariate Categorical Data. *Journal of the American Statistical Association*, 104(487):1042–1051.

Dwyer-Lindgren, L., Gakidou, E., Flaxman, A., and Wang, H. (2013). Error and bias in under-5 mortality estimates derived from birth histories with small sample sizes. *Population Health Metrics*, 11:1–17.

Egami, N., Fong, C. J., Grimmer, J., Roberts, M. E., and Stewart, B. M. (2022). How to make causal inferences using texts. *Science Advances*, 8(42):eabg2652.

Egami, N., Jacobs-Harukawa, M., Stewart, B. M., and Wei, H. (2023). Using large language model annotations for valid downstream statistical inference in social science: Design-based semi-supervised learning. *ArXiv preprint, abs/2306.04746*.

Erchick, D. J., Subedi, S., Verhulst, A., Guillot, M., Adair, L. S., Barros, A. J., Chasekwa, B., Christian, P., da Silva, B. G. C., Silveira, M. F., et al. (2023). Quality of vital event data for infant mortality estimation in prospective, population-based studies: an analysis of secondary data from asia, africa, and latin america. *Population health metrics*, 21(1):10.

Fabian, P. (2011). Scikit-learn: Machine learning in python. *Journal of machine learning research 12*, page 2825.

Fagbamigbe, A. F., Adeniji, F. I. P., and Morakinyo, O. M. (2022). Factors contributing to household wealth inequality in under-five deaths in low-and middle-income countries: decomposition analysis. *BMC public health*, 22(1):769.

Fahrmeir, L., Klasen, S., and Berger, U. (2002). Dynamic modelling of child mortality in developing countries: Application for zambia.

Fan, S., Liu, L., Perin, J., and McCormick, T. H. (2023). Bayesian age category reconciliation for age-and cause-specific under-five mortality estimates. *arXiv preprint arXiv:2302.11058*.

Fan, S., Visokay, A., Hoffman, K., Salerno, S., Liu, L., Leek, J. T., and McCormick, T. (2024). From narratives to numbers: Valid inference using language model predictions from verbal autopsies. In *First Conference on Language Modeling*.

Fenton, S. H., Stanfill, M. H., and Giannangelo, K. (2024). An icd for the digital world: What does the icd-11 research show? *MEDINFO 2023—The Future Is Accessible*, pages 58–62.

Fienberg, S. E. (1972). The analysis of incomplete multi-way contingency tables. *Biometrics*, pages 177–202.

Fienberg, S. E. (2006). Privacy and confidentiality in an e-commerce world: Data mining, data warehousing, matching and disclosure limitation. *Statistical Science*, pages 143–154.

Fienberg, S. E. and Holland, P. W. (1973). Simultaneous estimation of multinomial cell probabilities. *Journal of the American Statistical Association*, 68(343):683–691.

Fienberg, S. E. and Slavkovic, A. B. (2005). Preserving the confidentiality of categorical statistical data bases when releasing information for association rules. *Data Mining and Knowledge Discovery*, 11(2):155–180.

Fiksel, J., Datta, A., Amouzou, A., and Zeger, S. (2021). Generalized Bayes Quantification Learning under Dataset Shift. *Journal of the American Statistical Association*, pages 1–19.

Flaxman, A. D., Harman, L., Joseph, J., Brown, J., and Murray, C. J. (2018). A deidentified database of 11,979 verbal autopsy open-ended responses. *Gates open research*, 2:18.

Fottrell, E. and Byass, P. (2010). Verbal autopsy: methods in transition. *Epidemiologic reviews*, 32(1):38–55.

Fuchs, C. (1982). Maximum likelihood estimation and model selection in contingency tables with missing data. *Journal of the American Statistical Association*, 77(378):270–278.

Gao, M., Zhang, Z., Yu, G., Arık, S. Ö., Davis, L. S., and Pfister, T. (2020a). Consistency-Based Semi-supervised Active Learning: Towards Minimizing Labeling Cost. In Vedaldi, A., Bischof, H., Brox, T., and Frahm, J.-M., editors, *Computer Vision – ECCV 2020*, pages 510–526, Cham. Springer International Publishing.

Gao, X., Wang, D., Cai, Y., and Tu, D. (2020b). Cognitive Diagnostic Computerized Adaptive Testing for Polytomously Scored Items. *Journal of Classification*, 37(3):709–729.

Garenne, M. (2014). Prospects for automated diagnosis of verbal autopsies. *BMC medicine*, 12:1–5.

Gelfand, A. E. (2000). Gibbs sampling. *Journal of the American statistical Association*, 95(452):1300–1304.

Gelman, A., Hwang, J., and Vehtari, A. (2014). Understanding predictive information criteria for bayesian models. *Statistics and computing*, 24:997–1016.

Gelman, A. and Shalizi, C. R. (2013). Philosophy and the practice of bayesian statistics. *British Journal of Mathematical and Statistical Psychology*, 66(1):8–38.

Geman, S. and Geman, D. (1984). Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE transactions on pattern analysis and machine intelligence*, 6(6):721–741.

Gibbons, C., Richards, S., Valderas, J. M., Campbell, J., et al. (2017). Supervised machine learning algorithms can classify open-text feedback of doctor performance with human-level accuracy. *Journal of medical Internet research*, 19(3):e6533.

Gibbons, P. and Greenberg, E. (1989). Bayesian analysis of contingency tables with partially categorized data. *Typescript, Washington University, St. Louis, Missouri*, 63130(0825.62236):10–1080.

Gibbons, P. C. and Greenberg, E. (1994). Bayesian reconstruction of contingency tables with partially categorized data. *Communications in Statistics-Theory and Methods*, 23(12):3349–3359.

Gilbert, B., Fiksel, J., Wilson, E., Kalter, H., Kante, A., Akum, A., Blau, D., Bassat, Q., Macicame, I., Gudo, E. S., et al. (2023). Multi-cause calibration of verbal autopsy–based cause-specific mortality estimates of children and neonates in mozambique. *The American journal of tropical medicine and hygiene*, 108(5 Suppl):78.

Gilula, Z. and McCulloch, R. (2013). Multi level categorical data fusion using partially fused data. *Quantitative Marketing and Economics*, 11(3):353–377.

Global Burden of Disease Collaborative Network (2020). Global burden of disease study 2019 (gbd 2019) under-5 mortality by detailed age groups 1950-2019.

Glynn, R. J., Laird, N. M., and Rubin, D. B. (1993). Multiple imputation in mixture models for nonignorable nonresponse with follow-ups. *Journal of the American Statistical Association*, 88(423):984–993.

Golding, N., Burstein, R., Longbottom, J., Browne, A. J., Fullman, N., Osgood-Zimmerman, A., Earl, L., Bhatt, S., Cameron, E., Casey, D. C., et al. (2017). Mapping under-5 and neonatal mortality in africa, 2000–15: a baseline analysis for the sustainable development goals. *The Lancet*, 390(10108):2171–2182.

Grimmer, J. and Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political analysis*, 21(3):267–297.

Guillot, M., Romero Prieto, J., Verhulst, A., and Gerland, P. (2022). Modeling age patterns of under-5 mortality: Results from a log-quadratic model applied to high-quality vital registration data. *Demography*, 59(1):321–347.

Gunel, E. (1984). A bayesian analysis of the multinomial model for a dichotomous response with nonrespondents. *Communications in Statistics-Theory and Methods*, 13(6):737–751.

Guo, L. and Zheng, C. (2019). Termination rules for variable-length CD-CAT from the information theory perspective. *Frontiers in Psychology*, 10(112).

Haber, M., Chen, C. C., and David Williamson, C. (1991). Analysis of repeated categorical responses from fully and partially cross-classified data. *Communications in Statistics-Theory and Methods*, 20(10):3293–3313.

Hallett, T. B., Gregson, S., Kurwa, F., Garnett, G. P., Dube, S., Chawira, G., Mason, P. R., and Nyamukapa, C. A. (2010). Measuring and correcting biased child mortality statistics in countries with generalized epidemics of hiv infection. *Bulletin of the World Health Organization*, 88:761–768.

Hankin, R. K. (2010). A generalization of the dirichlet distribution. *Journal of Statistical Software*, 33:1–18.

Hastie, T. and Tibshirani, R. (1997). Classification by pairwise coupling. *Advances in neural information processing systems*, 10.

Hastie, T., Tibshirani, R., Friedman, J. H., and Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer.

He, C., Liu, L., Chu, Y., Perin, J., Dai, L., Li, X., Miao, L., Kang, L., Li, Q., Scherpbier, R., et al. (2017). National and subnational all-cause and cause-specific child mortality in china, 1996–2015: a systematic analysis with implications for the sustainable development goals. *The Lancet Global Health*, 5(2):e186–e197.

He, H., Bai, Y., Garcia, E. A., and Li, S. (2008). Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)*, pages 1322–1328. Ieee.

Hill, K., You, D., Inoue, M., and Oestergaard, M. (2012). Technical advisory group of united nations inter-agency group for child mortality estimation. child mortality estimation: accelerated progress in reducing global child mortality, 1990–2010. *PLoS Med*, 9(8):e1001303.

Hinga, A., Marsh, V., Nyaguara, A., Wamukoya, M., and Molyneux, S. (2021). The ethical implications of verbal autopsy: responding to emotional and moral distress. *BMC Medical Ethics*, 22(1):1–16.

Hirschberg, J. and Manning, C. D. (2015). Advances in natural language processing. *Science*, 349(6245):261–266.

Hocking, R. and Oxspring, H. (1974). The analysis of partially categorized contingency data. *Biometrics*, pages 469–483.

Hocking, R. R. and Oxspring, H. (1971). Maximum likelihood estimation with incomplete multinomial data. *Journal of the American Statistical Association*, 66(333):65–70.

Hoffman, K., Salerno, S., Afiaz, A., Leek, J. T., and McCormick, T. H. (2024). Do we really even need data? *arXiv preprint arXiv:2401.08702*.

Hoffman, M. D., Blei, D. M., Wang, C., and Paisley, J. (2013). Stochastic variational inference. *the Journal of machine Learning research*, 14(1):1303–1347.

Hoffman, M. D., Gelman, A., et al. (2014). The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. *J. Mach. Learn. Res.*, 15(1):1593–1623.

Hoi, S. C. H., Jin, R., and Lyu, M. R. (2006). Large-Scale Text Categorization by Batch Mode Active Learning. In *Proceedings of the 15th International Conference on World Wide Web*, WWW '06, pages 633–642, New York, NY, USA. Association for Computing Machinery.

Hornik, K., Leisch, F., Zeileis, A., and Plummer, M. (2003). Jags: A program for analysis of bayesian graphical models using gibbs sampling. In *Proceedings of DSC*, volume 2.

Hsu, C.-L., Wang, W.-C., and Chen, S.-Y. (2013). Variable-Length Computerized Adaptive Testing Based on Cognitive Diagnosis Models. *Applied Psychological Measurement*, 37(7):563–582.

Huebner, A. (2010). An overview of recent developments in cognitive diagnostic computer adaptive assessments. *Practical Assessment, Research, and Evaluation*, 15(1):3.

Huebner, A., Finkelman, M. D., and Weissman, A. (2018). Factors affecting the classification accuracy and average length of a variable-length cognitive diagnostic computerized test. *Journal of Computerized Adaptive Testing*, 6(1):1–14.

Ighodaro, A., Santner, T., and Brown, L. (1982). Admissibility and complete class results for the multinomial estimation problem with entropy and squared error loss. *Journal of Multivariate Analysis*, 12(4):469–479.

Ishwaran, H. and James, L. F. (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American statistical Association*, 96(453):161–173.

James, S. L., Flaxman, A. D., and Murray, C. J. (2011). Performance of the tariff method: validation of a simple additive algorithm for analysis of verbal autopsies. *Population Health Metrics*, 9:1–16.

Jaynes, E. T. (2003). *Probability theory: The logic of science*. Cambridge university press.

Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 186(1007):453–461.

Jiang, T. J. and Dickey, J. M. (2008). Bayesian methods for categorical data under informative censoring. *Bayesian Analysis*, 3(3):541–554.

Johnson, R. W. (1987). Simultaneous estimation of binomial n's. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 264–267.

Kadane, J. B. (1985). Is victimization chronic? a bayesian analysis of multinomial missing data. *Journal of econometrics*, 29(1-2):47–67.

Kalter, H. D., Salgado, R., Babille, M., Koffi, A. K., and Black, R. E. (2011). Social autopsy for maternal and child deaths: a comprehensive literature review to examine the concept and the development of the method. *Population health metrics*, 9:1–13.

Kaplan, M., de la Torre, J., and Barrada, J. (2015). New item selection methods for cognitive diagnosis computerized adaptive testing. *Applied Psychological Measurement*, 39(3):167–188.

Karson, M. and Wrobleski, W. (1970). A bayesian analysis of binomial data with a partially informative category. In *Proceedings of the Business and Economic Statistics Section, American Statistical Association*, pages 532–534.

Kateri, M. (2018). ϕ-divergence in contingency table analysis. *Entropy*, 20(5):324.

Kaufman, G. M. and King, B. (1973). A bayesian analysis of nonresponse in dichotomous processes. *Journal of the American Statistical Association*, 68(343):670–678.

Kaul, T., Kellerhuis, B. E., Damen, J. A., Schuit, E., Jenniskens, K., van Smeden, M., Reitsma, J. B., Hooft, L., Moons, K. G., and Yang, B. (2024). Methodological quality assessment tools for diagnosis and prognosis research: overview and guidance. *Journal of Clinical Epidemiology*, page 111609.

Kaushal, V., Iyer, R., Kothawade, S., Mahadev, R., Doctor, K., and Ramakrishnan, G. (2019). Learning from less data: A unified data subset selection and active learning framework for computer vision. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1289–1299. IEEE.

King, G. and Zeng, L. (2001). Logistic regression in rare events data. *Political analysis*, 9(2):137–163.

Kleinberg, J., Ludwig, J., Mullainathan, S., and Obermeyer, Z. (2015). Prediction policy problems. *American Economic Review*, 105(5):491–495.

Klir, G. and Wierman, M. (1999). *Uncertainty-based information: elements of generalized information theory*, volume 15. Springer Science & Business Media.

Kneib, T. (2015). Applied statistical inference: Likelihood and bayes. l. held and d. sabanés bové (2014). heidelberg: Springer. 376 pages, isbn: 3642378862.

Knox, D., Lucas, C., and Cho, W. K. T. (2022). Testing causal theories with learned proxies. *Annual Review of Political Science*, 25(1):419–441.

Koch, G. G., Imrey, P. B., and Reinfurt, D. W. (1972). Linear model analysis of categorical data with incomplete response vectors. *Biometrics*, pages 663–692.

Korenromp, E. L., Williams, B. G., Gouws, E., Dye, C., and Snow, R. W. (2003). Measurement of trends in childhood malaria mortality in africa: an assessment of progress toward targets based on verbal autopsy. *The Lancet infectious diseases*, 3(6):349–358.

Kosmopoulos, A., Partalas, I., Gaussier, E., Paliouras, G., and Androutsopoulos, I. (2015). Evaluation measures for hierarchical classification: A unified view and novel approaches. *Data Mining and Knowledge Discovery*, 29(3):820–865.

Kramer, O. and Kramer, O. (2016). Scikit-learn. *Machine learning for evolution strategies*, pages 45–53.

Kullback, S. (1997). *Information theory and statistics*. Courier Corporation.

Kumar, P. and Gupta, A. (2020). Active Learning Query Strategies for Classification, Regression, and Clustering: A Survey. *Journal of Computer Science and Technology*, 35(4):913–945.

Kunihama, T., Li, Z. R., Clark, S. J., and McCormick, T. H. (2020a). Bayesian factor models for probabilistic cause of death assessment with verbal autopsies. *The Annals of Applied Statistics*, 14(1):241–256.

Kunihama, T., Li, Z. R., Clark, S. J., and McCormick, T. H. (2020b). Bayesian factor models for probabilistic cause of death assessment with verbal autopsies. *The annals of applied statistics*, 14(1):241.

Kuss, M., Jäkel, F., and Wichmann, F. A. (2005). Bayesian inference for psychometric functions. *Journal of Vision*, 5(5):8–8.

Lai, V. D., Ngo, N. T., Veyseh, A. P. B., Man, H., Dernoncourt, F., Bui, T., and Nguyen, T. H. (2023). Chatgpt beyond english: Towards a comprehensive evaluation of large language models in multilingual learning. *arXiv preprint arXiv:2304.05613*.

Lawn, J. E., Blencowe, H., Oza, S., You, D., Lee, A. C., Waiswa, P., Lalli, M., Bhutta, Z., Barros, A. J., Christian, P., et al. (2014). Every newborn: progress, priorities, and potential beyond survival. *The lancet*, 384(9938):189–205.

Lawn, J. E., Blencowe, H., Waiswa, P., Amouzou, A., Mathers, C., Hogan, D., Flenady, V., Frøen, J. F., Qureshi, Z. U., Calderwood, C., et al. (2016). Stillbirths: rates, risk factors, and acceleration towards 2030. *The Lancet*, 387(10018):587–603.

Lawrence, E., Liu, C., Wiel, S. V., and Zhang, J. (2009). A new method of multinomial inference using dempster-shafer theory. *Unpublished Manuscript*.

Lawson, A. B., Banerjee, S., Haining, R. P., and Ugarte, M. D. (2016). *Handbook of spatial epidemiology*. CRC press.

Lemoine, N. P. (2019). Moving beyond noninformative priors: why and how to choose weakly informative priors in bayesian analyses. *Oikos*, 128(7):912–928.

Leonard, T. (1977). A bayesian approach to some multinomial estimation and pretesting problems. *Journal of the American Statistical Association*, 72(360a):869–874.

Lewis, D. D. and Gale, W. A. (1994a). A sequential algorithm for training text classifiers. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 3–12.

Lewis, D. D. and Gale, W. A. (1994b). A Sequential Algorithm for Training Text Classifiers. In *Proceedings of the 17th Annual International ACM–SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '94, pages 3–12, London. Springer-Verlag.

Li, Y. and Oliva, J. (2021). Active feature acquisition with generative surrogate models. In *International Conference on Machine Learning*, pages 6450–6459. PMLR.

Li, Z. R., McComick, T. H., and Clark, S. J. (2019). Using bayesian latent gaussian graphical models to infer symptom associations in verbal autopsies. *Bayesian analysis*, 15(3):781.

Li, Z. R., McCormick, T. H., and Clark, S. J. (2020). Using Bayesian latent Gaussian graphical models to infer symptom associations in verbal autopsies. *Bayesian Analysis*, 15(3):781.

Li, Z. R., Thomas, J., Choi, E., McCormick, T. H., and Clark, S. J. (2022). The openva toolkit for verbal autopsies. *arXiv preprint arXiv:2109.0824*.

Li, Z. R., Thomas, J., Choi, E., McCormick, T. H., and Clark, S. J. (2023). The openva toolkit for verbal autopsies. *The R journal*, 14(4):316.

Li, Z. R., Wu, Z., Chen, I., and Clark, S. J. (2021). Bayesian nested latent class models for cause-of-death assignment using verbal autopsies across multiple domains. *arXiv preprint arXiv:2112.12186*.

Liu, J., Ying, Z., and Zhang, S. (2015). A rate function approach to computerized adaptive testing for cognitive diagnosis. *Psychometrika*, 80(2):468–490.

Liu, J. S. and Wu, Y. N. (1999). Parameter expansion for data augmentation. *Journal of the American Statistical Association*, 94(448):1264–1274.

Liu, L., Oza, S., Hogan, D., Chu, Y., Perin, J., Zhu, J., Lawn, J. E., Cousens, S., Mathers, C., and Black, R. E. (2016). Global, regional, and national causes of under-5 mortality in 2000–15: an updated systematic analysis with implications for the sustainable development goals. *The Lancet*, 388(10063):3027–3035.

Liu, R., Greenstein, J. L., Sarma, S. V., and Winslow, R. L. (2019). Natural language processing of clinical notes for improved early prediction of septic shock in the icu. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 6103–6108. IEEE.

Loh, P., Fottrell, E., Beard, J., Bar-Zeev, N., Phiri, T., Banda, M., Makwenda, C., Bird, J., and King, C. (2021). Added value of an open narrative in verbal autopsies: a mixed-methods evaluation from malawi. *BMJ Paediatrics Open*, 5(1).

Lord, F. M. (1971). Robbins-monro procedures for tailored testing. *Educational and Psychological Measurement*, 31(1):3–31.

Madley-Dowd, P., Hughes, R., Tilling, K., and Heron, J. (2019). The proportion of missing data should not be used to guide decisions on multiple imputation. *Journal of Clinical Epidemiology*, 110:63–73.

Manaka, T., van Zyl, T., and Kar, D. (2022). Improving cause-of-death classification from verbal autopsy reports. In *Southern African Conference for Artificial Intelligence Research*, pages 46–59. Springer.

Manski, C. F. (2009). *Identification for prediction and decision*. Harvard University Press.

Marquez, N. and Wakefield, J. (2021). Harmonizing child mortality data at disparate geographic levels. *Statistical methods in medical research*, 30(5):1187–1210.

Martinez-Beneito, M. A., Botella-Rocamora, P., and Banerjee, S. (2016). Towards a multidimensional approach to bayesian disease mapping. *Bayesian analysis*, 12(1):239.

McCallum, A. and Nigam, K. (1998). Employing EM and Pool-Based Active Learning for Text Classification. In *Proceedings of the 15th International Conference on Machine Learning*, ICML '98, pages 350–358, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

McCormick, T. H., Li, Z. R., Calvert, C., Crampin, A. C., Kahn, K., and Clark, S. J. (2016a). Probabilistic Cause-of-Death Assignment Using Verbal Autopsies. *Journal of the American Statistical Association*, 111(515):1036–1049.

McCormick, T. H., Li, Z. R., Calvert, C., Crampin, A. C., Kahn, K., and Clark, S. J. (2016b). Probabilistic cause-of-death assignment using verbal autopsies. *Journal of the American Statistical Association*, 111(515):1036–1049.

McLachlan, G. J., Lee, S. X., and Rathnayake, S. I. (2019). Finite mixture models. *Annual review of statistics and its application*, 6(1):355–378.

Mejía-Guevara, I., Zuo, W., Bendavid, E., Li, N., and Tuljapurkar, S. (2019). Age distribution, trends, and forecasts of under-5 mortality in 31 sub-saharan african countries: A modeling study. *PLoS medicine*, 16(3):e1002757.

Melville, P., Saar-Tsechansky, M., Provost, F., and Mooney, R. (2004). Active feature-value acquisition for classifier induction. In *Fourth IEEE International Conference on Data Mining (ICDM'04)*, pages 483–486. IEEE.

Memon, Z., Fridman, D., Soofi, S., Ahmed, W., Muhammad, S., Rizvi, A., Ahmed, I., Wright, J., Cousens, S., and Bhutta, Z. A. (2023). Predictors and disparities in neonatal and under 5 mortality in rural pakistan: cross sectional analysis. *The Lancet Regional Health-Southeast Asia*, 15.

Miao, J., Miao, X., Wu, Y., Zhao, J., and Lu, Q. (2023). Assumption-lean and data-adaptive post-prediction inference. *arXiv preprint arXiv:2311.14220*.

Mikkelsen, L., Phillips, D. E., AbouZahr, C., Setel, P. W., De Savigny, D., Lozano, R., and Lopez, A. D. (2015). A global assessment of civil registration and vital statistics systems: monitoring data quality and progress. *The Lancet*, 386(10001):1395–1406.

Millstein, F. (2020). *Natural language processing with python: natural language processing using NLTK*. Frank Millstein.

Molenberghs, G., Fitzmaurice, G., Kenward, M. G., Tsiatis, A., and Verbeke, G. (2014). *Handbook of missing data methodology*. CRC Press.

Moran, K. R., Turner, E. L., Dunson, D., and Herring, A. H. (2021). Bayesian hierarchical factor regression models to infer cause of death from verbal autopsy data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 70(3):532–557.

Mortier, T., Wydmuch, M., Dembczyński, K., Hüllermeier, E., and Waegeman, W. (2021). Efficient set-valued prediction in multi-class classification. *Data Mining and Knowledge Discovery*, 35(4):1435–1469.

Mosley, W. H. and Chen, L. C. (2003). An analythical framework for the study of child survival in developing countries. *Bulletin of the world Health Organization*, 81:140–145.

Mukaka, M., White, S. A., Terlouw, D. J., Mwapasa, V., Kalilani-Phiri, L., and Faragher, E. B. (2016). Is using multiple imputation better than complete case analysis for estimating a prevalence (risk) difference in randomized controlled trials when binary outcome observations are missing? *Trials*, 17:1–12.

Mulick, A. R., Oza, S., Prieto-Merino, D., Villavicencio, F., Cousens, S., and Perin, J. (2021). A bayesian hierarchical model with integrated covariate selection and misclassification matrices to estimate neonatal and child causes of death. *medRxiv*, pages 2021–02.

Mulick, A. R., Oza, S., Prieto-Merino, D., Villavicencio, F., Cousens, S., and Perin, J. (2022). A bayesian hierarchical model with integrated covariate selection and misclassification matrices to estimate neonatal and child causes of death. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 185(4):2097–2120.

Mullainathan, S. and Spiess, J. (2017). Machine learning: an applied econometric approach. *Journal of Economic Perspectives*, 31(2):87–106.

Murray, C. J., Aravkin, A. Y., Zheng, P., Abbafati, C., Abbas, K. M., Abbasi-Kangevari, M., Abd-Allah, F., Abdelalim, A., Abdollahi, M., Abdollahpour, I., et al. (2020). Global burden of 87 risk factors in 204 countries and territories, 1990–2019: a systematic analysis for the global burden of disease study 2019. *The lancet*, 396(10258):1223–1249.

Murray, C. J., Laakso, T., Shibuya, K., Hill, K., and Lopez, A. D. (2007). Can we achieve millennium development goal 4? new analysis of country trends and forecasts of under-5 mortality to 2015. *The lancet*, 370(9592):1040–1054.

Murray, C. J., Lopez, A. D., Black, R., Ahuja, R., Ali, S. M., Baqui, A., Dandona, L., Dantzer, E., Das, V., Dhingra, U., et al. (2011a). Population health metrics research consortium gold standard verbal autopsy validation study: design, implementation, and development of analysis datasets. *Population health metrics*, 9(1):27.

Murray, C. J., Lopez, A. D., Shibuya, K., and Lozano, R. (2011b). Verbal autopsy: advancing science, facilitating application.

Murray, C. J., Lozano, R., Flaxman, A. D., Serina, P., Phillips, D., Stewart, A., James, S. L., Vahdatpour, A., Atkinson, C., Freeman, M. K., et al. (2014). Using verbal autopsy to measure causes of death: the comparative performance of existing methods. *BMC medicine*, 12:1–19.

Naghavi, M., Makela, S., Foreman, K., O'Brien, J., Pourmalek, F., and Lozano, R. (2010). Algorithms for enhancing public health utility of national causes-of-death data. *Population health metrics*, 8:1–14.

Naradowsky, J., Riedel, S., and Smith, D. A. (2012). Improving nlp through marginalization of hidden syntactic structure. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 810–820.

National Institute of Population Research and Training (NIPORT), Mitra and Associates, and ICF International (2020). Bangladesh demographic and health survey 2017-18.

Navigli, R., Conia, S., and Ross, B. (2023). Biases in large language models: origins, inventory, and discussion. *ACM Journal of Data and Information Quality*, 15(2):1–21.

Nayak, T. K. and Naik, D. N. (1989). Estimating multinomial cell probabilities under quadratic loss. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 38(1):3–10.

Ng, K. W., Tian, G.-L., and Tang, M.-L. (2011). Dirichlet and related distributions: Theory, methods and applications.

Nichols, E., Pettrone, K., Vickers, B., Gebrehiwet, H., Surek-Clark, C., Leitao, J., Amouzou, A., Blau, D. M., Bradshaw, D., Abdelilah, E. M., Groenewald, P., Munkombwe, B., Mwango, C., Notzon, F. S., Odhiambo, S. B., and Scanlon, P. (2022). Mixed-methods analysis of select issues reported in the 2016 world health organization verbal autopsy questionnaire. *Plos one*, 17(10):e0274304.

Nichols, E. K., Byass, P., Chandramohan, D., Clark, S. J., Flaxman, A. D., Jakob, R., Leitao, J., Maire, N., Rao, C., Riley, I., et al. (2018). The who 2016 verbal autopsy instrument: An international standard suitable for automated analysis by interva, insilicova, and tariff 2.0. *PLoS medicine*, 15(1):e1002486.

Nijman, S. W., Leeuwenberg, A., Beekers, I., Verkouter, I., Jacobs, J., Bots, M., Asselbergs, F., Moons, K. G., and Debray, T. P. (2022). Missing data is poorly handled and reported in prediction model studies using machine learning: a literature review. *Journal of clinical epidemiology*, 142:218–229.

NIPORT (2012). *Bangladesh Demographic and Health Survey 2011: Preliminary Report*. NIPORT.

NIPORT and ICF (2016). Bangladesh health facility survey 2017.

of Population Research, N. I., (NIPOR), T., and ICF. *Bangladesh Demographic and Health Survey 2017-18*. NIPORT and ICF, Dhaka, Bangladesh, and Rockville, Maryland, USA.

of Population Research, N. I., Training NIPORT/Bangladesh, M., Associates/Bangladesh, and International, I. *Bangladesh Demographic and Health Survey 2011*. NIPORT, Mitra and Associates, and ICF International, Dhaka, Bangladesh.

Ogburn, E. L., Rudolph, K. E., Morello-Frosch, R., Khan, A., and Casey, J. A. (2021). A warning about using predicted values from regression models for epidemiologic inquiry. *American Journal of Epidemiology*, 190(6):1142–1147.

Oxspring, H. and LaMotte, L. (1977). The design of incomplete multinomial samples: the nested case. *Technometrics*, 19(1):59–63.

Paulino, C. D. M. and de Braganca Pereira, C. A. (1995). Bayesian methods for categorical data under informative general censoring. *Biometrika*, 82(2):439–446.

Pearl, J. and Bareinboim, E. (2014). External validity: From do-calculus to transportability across populations. *Statistical Science*, 29(4):579–595.

Perin, J., Chu, Y., Villaviciencio, F., Schumacher, A., McCormick, T., Guillot, M., and Liu, L. (2022a). Adapting and validating the log quadratic model to derive under-five age-and cause-specific mortality (u5acsm): a preliminary analysis. *Population health metrics*, 20:1–12.

Perin, J., Mulick, A., Yeung, D., Villavicencio, F., Lopez, G., Strong, K. L., Prieto-Merino, D., Cousens, S., Black, R. E., and Liu, L. (2022b). Global, regional, and national causes of under-5 mortality in 2000–19: an updated systematic analysis with implications for the sustainable development goals. *The Lancet Child & Adolescent Health*, 6(2):106–115.

Peskoff, D. and Stewart, B. M. (2023). Credible without credit: Domain experts assess generative language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 427–438.

Peskov, D. and Hangya, V. (2021). Adapting entities across languages and cultures. *Findings of the Association for Computational Linguistics: EMNLP 2021*.

Peter, J., Barron, P., and Pillay, Y. (2016). Using mobile technology to improve maternal, child and youth health and treatment of hiv patients. *SAMJ: South African Medical Journal*, 106(1):3–4.

Phyo, A. Z. Z., Freak-Poli, R., Craig, H., Gasevic, D., Stocks, N. P., Gonzalez-Chica, D. A., and Ryan, J. (2020). Quality of life and mortality in the general population: a systematic review and meta-analysis. *BMC public health*, 20:1–20.

Rajkomar, A., Oren, E., Chen, K., Dai, A. M., Hajaj, N., Hardt, M., Liu, P. J., Liu, X., Marcus, J., Sun, M., et al. (2018). Scalable and accurate deep learning with electronic health records. *NPJ digital medicine*, 1(1):18.

Rao, J. N. and Molina, I. (2015). *Small area estimation*. John Wiley & Sons.

Rasmussen, C. E. (2003). Gaussian processes in machine learning. In *Summer school on machine learning*, pages 63–71. Springer.

Richardson, S. and Green, P. J. (1997). On bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 59(4):731–792.

Roberts, G. O. and Rosenthal, J. S. (2004). General state space markov chains and mcmc algorithms.

Roberts, G. O. and Rosenthal, J. S. (2009). Examples of adaptive mcmc. *Journal of computational and graphical statistics*, 18(2):349–367.

Rodgers, W. L. (1984). An evaluation of statistical matching. *Journal of Business & Economic Statistics*, 2(1):91–102.

Roth, G. A., Abate, D., Abate, K. H., Abay, S. M., Abbafati, C., Abbasi, N., Abbastabar, H., Abd-Allah, F., Abdela, J., Abdelalim, A., et al. (2018). Global, regional, and national age-sex-specific mortality for 282 causes of death in 195 countries and territories, 1980–2017: a systematic analysis for the global burden of disease study 2017. *The lancet*, 392(10159):1736–1788.

Roy, N. and McCallum, A. (2001). Toward optimal active learning through monte carlo estimation of error reduction. *ICML, Williamstown*, 2:441–448.

Rubin, D. B. (1974). Characterizing the estimation of parameters in incomplete-data problems. *Journal of the American Statistical Association*, 69(346):467–474.

Rubin, D. B. (1986). Statistical matching using file concatenation with adjusted weights and multiple imputations. *Journal of Business & Economic Statistics*, 4(1):87–94.

Salakhutdinov, R. and Mnih, A. (2008). Bayesian Probabilistic Matrix Factorization Using Markov Chain Monte Carlo. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, pages 880–887, New York, NY, USA. Association for Computing Machinery.

Salvatier, J., Wiecki, T. V., and Fonnesbeck, C. (2016). Probabilistic programming in python using pymc3. *PeerJ Computer Science*, 2:e55.

Sanghera, S., Coast, J., Martin, R. M., Donovan, J. L., and Mohiuddin, S. (2018). Cost-effectiveness of prostate cancer screening: a systematic review of decision-analytical models. *BMC cancer*, 18:1–15.

Sankoh, O. and Byass, P. (2014). Time for civil registration with verbal autopsy. *The Lancet Global Health*, 2(12):e693–e694.

Schmid, C. H., Trikalinos, T. A., and Olkin, I. (2014). Bayesian network meta-analysis for unordered categorical outcomes with incomplete data. *Research Synthesis Methods*, 5(2):162–185.

Schumacher, A. E., McCormick, T. H., Wakefield, J., Chu, Y., Perin, J., Villavicencio, F., Simon, N., and Liu, L. (2020). A flexible bayesian framework to estimate age-and cause-specific child mortality over time from sample registration data. *arXiv preprint arXiv:2003.00401*.

Schumacher, A. E., McCormick, T. H., Wakefield, J., Chu, Y., Perin, J., Villavicencio, F., Simon, N., and Liu, L. (2022). A flexible bayesian framework to estimate age-and cause-specific child mortality over time from sample registration data. *The annals of applied statistics*, 16(1):124.

Seeger, M. (2004). Gaussian processes for machine learning. *International journal of neural systems*, 14(02):69–106.

Sentz, B. and Ferson, S. (2002a). Combination of Evidence in DempsterShafer Theory. Technical report, Sandia National Laboratory.

Sentz, K. and Ferson, S. (2002b). Combination of evidence in dempster-shafer theory.

Sepanlou, S. G., Aliabadi, H. R., Malekzadeh, R., Naghavi, M., in Middle East Collaborators, G. C. M., et al. (2022). Neonate, infant, and child mortality in north africa and middle east by cause: An analysis for the global burden of disease study 2019. *Archives of Iranian medicine*, 25(12):767.

Serina, P., Riley, I., Stewart, A., Flaxman, A. D., Lozano, R., Mooney, M. D., Luning, R., Hernandez, B., Black, R., Ahuja, R., et al. (2015a). A shortened verbal autopsy instrument for use in routine mortality surveillance systems. *BMC Medicine*, 13:1–10.

Serina, P., Riley, I., Stewart, A., James, S. L., Flaxman, A. D., Lozano, R., Hernandez, B., Mooney, M. D., Luning, R., Black, R., et al. (2015b). Improving performance of the Tariff method for assigning causes of death to verbal autopsies. *BMC Medicine*, 13(1):1.

Sethi, R., Dey, D., and Nath, D. (2008). Sample registration system: Information on health status. *Demography India*, 37:51–65.

Settles, B. (2009). Active Learning Literature Survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison.

Settles, B. (2011). From theories to queries: Active learning in practice. In *Active learning and experimental design workshop in conjunction with AISTATS 2010*, pages 1–18. JMLR Workshop and Conference Proceedings.

Seung, H. S., Opper, M., and Sompolinsky, H. (1992). Query by committee. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 287–294.

Sharrow, D., Hug, L., You, D., Alkema, L., Black, R., Cousens, S., Croft, T., Gaigbe-Togbe, V., Gerland, P., Guillot, M., et al. (2022). Global, regional, and national trends in under-5 mortality between 1990 and 2019 with scenario-based projections until 2030: a systematic analysis by the un inter-agency group for child mortality estimation. *The Lancet Global Health*, 10(2):e195–e206.

Shaw, J. and Sekalala, S. (2023). Health data justice: building new norms for health data governance. *NPJ Digital Medicine*, 6(1):30.

Shawon, M. T. H., Ashrafi, S. A. A., Azad, A. K., Firth, S. M., Chowdhury, H., Mswia, R. G., Adair, T., Riley, I., Abouzahr, C., and Lopez, A. D. (2021). Routine mortality surveillance to identify the cause of death pattern for out-of-hospital adult (aged 12+ years) deaths in bangladesh: introduction of automated verbal autopsy. *BMC public health*, 21:1–11.

Shefrin, H. (1981). On the combinatorial structure of bayesian learning with imperfect information. *Discrete Mathematics*, 37(2-3):245–254.

Shefrin, H. (1983). Markov chains, imperfect state information, and bayesian learning. *Mathematical Modelling*, 4(1):1–7.

Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., et al. (2023). Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Smith, P. J., Choi, S. C., and Gunel, E. (1985). Bayesian analysis of a 2× 2 contingency table with both completely and partially cross-classified data. *Journal of Educational Statistics*, 10(1):31–43.

Smith, P. J. and Gunel, E. (1984). Practical bayesian approaches to the analystcs of a 2× 2 contingency table with incompletely categorized data. *Communications in Statistics-Theory and Methods*, 13(16):1941–1963.

Soleman, N., Chandramohan, D., and Shibuya, K. (2006). Verbal autopsy: current practices and challenges. *Bulletin of the World Health Organization*, 84(3):239–245.

Song, P., Theodoratou, E., Li, X., Liu, L., Chu, Y., Black, R. E., Campbell, H., Rudan, I., and Chan, K. Y. (2016). Causes of death in children younger than five years in china in 2015: an updated analysis. *Journal of global health*, 6(2):020802.

Streatfield, P. K., Khan, W. A., Bhuiya, A., Hanifi, S. M., Alam, N., Diboulo, E., Niamba, L., Sié, A., Lankoande, B., Millogo, R., et al. (2014). Mortality from external causes in africa and asia: evidence from indepth health and demographic surveillance system sites. *Global health action*, 7(1):25366.

Sundberg, R. (1974). Maximum likelihood theory for incomplete data from an exponential family. *Scandinavian Journal of Statistics*, pages 49–58.

Surek-Clark, C. (2020). Verbal autopsy interview standardization study. *The Anthropological Demography of Health*, page 301.

Tan, Y. V. and Roy, J. (2019). Bayesian additive regression trees and the general bart model. *Statistics in medicine*, 38(25):5048–5069.

Tanner, M. A. and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American statistical Association*, 82(398):528–540.

Tatsuoka, C. (2002a). Data analytic methods for latent partially ordered classification models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 51(3):337–350.

Tatsuoka, C. (2002b). Data analytic methods for latent partially ordered classification models. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 51(3):337–350.

Tatsuoka, C. and Ferguson, T. (2003). Sequential classification on partially ordered sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(1):143–157.

Teh, Y. W. et al. (2010). Dirichlet process. *Encyclopedia of machine learning*, 1063:280–287.

Thomas, L.-M., D'Ambruoso, L., and Balabanova, D. (2018). Verbal autopsy in health policy and systems: a literature review. *BMJ global health*, 3(2):e000639.

Tian, G.-L., Ng, K. W., and Geng, Z. (2003). Bayesian computation for contingency tables with incomplete cell-counts. *Statistica Sinica*, pages 189–206.

Tomek, I. (1976). Two modifications of cnn.

Tong, S. and Koller, D. (2001). Support Vector Machine Active Learning With Applications To Text Classification. *The Journal of Machine Learning Research*, 2:45–66.

Turnbull, B. W. (1976). The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 38(3):290–295.

Uneke, C. J., Uro-Chukwu, H. C., and Chukwu, O. E. (2019). Validation of verbal autopsy methods for assessment of child mortality in sub-saharan africa and the policy implication: a rapid review. *Pan African Medical Journal*, 33(1).

United Nations Inter-Agency Group For Child Mortality Estimation (2020). Levels & trends in child mortality 2020 technical report. *United Nations Children's Fund*.

Utazi, C. E., Thorley, J., Alegana, V. A., Ferrari, M. J., Takahashi, S., Metcalf, C. J. E., Lessler, J., and Tatem, A. J. (2018). High resolution age-structured mapping of childhood vaccination coverage in low and middle income countries. *Vaccine*, 36(12):1583–1591.

Van der Linden, W. J. and Glas, C. A. (2010). *Elements of adaptive testing*, volume 10. Springer.

Van Malderen, C., Amouzou, A., Barros, A. J., Masquelier, B., Van Oyen, H., and Speybroeck, N. (2019). Socioeconomic factors contributing to under-five mortality in sub-saharan africa: a decomposition analysis. *BMC public health*, 19:1–19.

Victora, C. G., Requejo, J. H., Barros, A. J., Berman, P., Bhutta, Z., Boerma, T., Chopra, M., De Francisco, A., Daelmans, B., Hazel, E., et al. (2016). Countdown to 2015: a decade of tracking progress for maternal, newborn, and child survival. *The Lancet*, 387(10032):2049–2059.

Victora, C. G., Wagstaff, A., Schellenberg, J. A., Gwatkin, D., Claeson, M., and Habicht, J.-P. (2003). Applying an equity lens to child health and mortality: more of the same is not enough. *The Lancet*, 362(9379):233–241.

Villavicencio, F., Perin, J., Eilerts-Spinelli, H., Yeung, D., Prieto-Merino, D., Hug, L., Sharrow, D., You, D., Strong, K. L., Black, R. E., et al. (2024). Global, regional, and national causes of death in children and adolescents younger than 20 years: an open data portal with estimates for 2000–21. *The Lancet Global Health*, 12(1):e16–e17.

Wahab, A., Choiriyyah, I., and Wilopo, S. A. (2017). Determining the cause of death: mortality surveillance using verbal autopsy in indonesia. *The American Journal of Tropical Medicine and Hygiene*, 97(5):1461.

Wakefield, J. (2007). Disease mapping and spatial regression with count data. *Biostatistics*, 8(2):158–183.

Wakefield, J. et al. (2013). *Bayesian and frequentist regression methods*, volume 23. Springer.

Walker, S. (1996). A bayesian maximum a posteriori algorithm for categorical data under informative general censoring. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 45(3):293–298.

Walley, P. (1996). Inferences from multinomial data: learning about a bag of marbles. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1):3–34.

Wang, C. (2013a). Mutual information item selection method in cognitive diagnostic computerized adaptive testing with short test length. *Educational and Psychological Measurement*, 73(6):1017–1035.

Wang, C. (2013b). Mutual Information Item Selection Method in Cognitive Diagnostic Computerized Adaptive Testing With Short Test Length. *Educational and Psychological Measurement*, 73(6):1017–1035.

Wang, C., Chang, H.-H., and Huebner, A. (2011). Restrictive stochastic item selection methods in cognitive diagnostic computerized adaptive testing. *Journal of Educational Measurement*, 48(3):255–273.

Wang, L., Yang, G., Jiemin, M., Rao, C., Wan, X., Dubrovsky, G., and Lopez, A. D. (2007). Evaluation of the quality of cause of death statistics in rural china using verbal autopsies. *Journal of Epidemiology & Community Health*, 61(6):519–526.

Wang, S., McCormick, T. H., and Leek, J. T. (2020). Methods for correcting inference based on outcomes predicted by machine learning. *Proceedings of the National Academy of Sciences*, 117(48):30266–30275.

Weiss, D. J. and Kingsbury, G. G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement*, 21(4):361–375.

Weiss, G. M. and Provost, F. (2003). Learning when training data are costly: The effect of class distribution on tree induction. *Journal of artificial intelligence research*, 19:315–354.

Weissman, A. (2007). Mutual Information Item Selection in Adaptive Classification Testing. *Educational and Psychological Measurement*, 67(1):41–58.

Wells, B. J., Chagin, K. M., Nowacki, A. S., and Kattan, M. W. (2013). Strategies for handling missing data in electronic health record derived data. *Egems*, 1(3):1035.

White, I. R., Royston, P., and Wood, A. M. (2011). Multiple imputation using chained equations: issues and guidance for practice. *Statistics in medicine*, 30(4):377–399.

Williamson, G. D. and Haber, M. (1994). Models for three-dimensional contingency tables with completely and partially cross-classified data. *Biometrics*, pages 194–203.

Woolson, R. F. and Clarke, W. R. (1984). Analysis of categorical incomplete longitudinal data. *Journal of the Royal Statistical Society: Series A (General)*, 147(1):87–99.

World Health Organization (2017). Verbal autopsy standards: The 2022 who verbal autopsy instrument. Technical report, World Health Organization, Geneva. License: CC BY-NC-SA 3.0 IGO.

World Health Organization (2021a). WHO civil registration and vital statistics strategic implementation plan 2021-2025.

World Health Organization (2021b). Who civil registration and vital statistics strategic implementation plan 2021-2025. Technical report, World Health Organization. License: CC BY-NC-SA 3.0 IGO.

World Health Organization (2022a). Icd-11 2022 release. WHO.

World Health Organization (2022b). Revision of the 2016 WHO verbal autopsy instrument.

World Health Organization (2022c). Verbal autopsy standards: verbal autopsy field interviewer manual for the 2022 WHO verbal autopsy instrument.

Wu, Z., Li, Z. R., Chen, I., and Li, M. (2021a). Tree-informed bayesian multi-source domain adaptation: cross-population probabilistic cause-of-death assignment using verbal autopsy. *arXiv preprint arXiv:2112.10978*.

Wu, Z., Li, Z. R., Chen, I., and Li, M. (2021b). Tree-informed Bayesian multi-source domain adaptation: cross-population probabilistic cause-of-death assignment using verbal autopsy. *arXiv preprint arXiv:2112.10978*.

Wyber, R., Vaillancourt, S., Perry, W., Mannava, P., Folaranmi, T., and Celi, L. A. (2015). Big data in global health: improving health in low-and middle-income countries. *Bulletin of the World Health Organization*, 93:203–208.

Xu, X., Chang, H., and Douglas, J. (2003). A simulation study to compare CAT strategies for cognitive diagnosis. In *annual meeting of the American Educational Research Association, Chicago*, page 34.

Xu, X. and Douglas, J. (2006). Computerized adaptive testing under nonparametric IRT models. *Psychometrika*, 71(1):121–137.

Ye, Z., Liu, P., Fu, J., and Neubig, G. (2021). Towards more fine-grained and reliable nlp performance prediction. *arXiv preprint arXiv:2102.05486*.

Yilema, S. A., Shiferaw, Y. A., Zewotir, T. T., and Muluneh, E. K. (2022). Spatial small area estimates of undernutrition for under five children in ethiopia via combining survey and census data. *Spatial and Spatio-temporal Epidemiology*, 42:100509.

Yoo, D. and Kweon, I. S. (2019). Learning loss for active learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 93–102.

Yoshida, T., Fan, T. S., McCormick, T., Zhenke, W., and Li, Z. R. (2023). Bayesian active questionnaire design for cause-of-death assignment using verbal autopsies. In *Conference on Health, Inference, and Learning*, pages 37–49. PMLR.

You, D., Hug, L., Ejdemyr, S., Idele, P., Hogan, D., Mathers, C., Gerland, P., New, J. R., and Alkema, L. (2015). Global, regional, and national levels and trends in under-5 mortality between 1990 and 2015, with scenario-based projections to 2030: a systematic analysis by the un inter-agency group for child mortality estimation. *The Lancet*, 386(10010):2275–2286.

Zadeh, L. A. (1986). A simple view of the dempster-shafer theory of evidence and its implication for the rule of combination. *AI magazine*, 7(2):85–85.

Zamboni, K., Singh, S., Tyagi, M., Hill, Z., Hanson, C., and Schellenberg, J. (2021). Effect of collaborative quality improvement on stillbirths, neonatal mortality and newborn care practices in hospitals of telangana and andhra pradesh, india: evidence from a quasi-experimental mixed-methods study. *Implementation Science*, 16:1–17.

Zhang, C., Taylor, S. J., Cobb, C., and Sekhon, J. (2020). Active matrix factorization for surveys. *The Annals of Applied Statistics*, 14(3):1182 – 1206.

Zhang, C. and Yin, G. (2019). Fast and stable maximum likelihood estimation for incomplete multinomial models. In *International Conference on Machine Learning*, pages 7463–7471. PMLR.

Zhao, J., Shang, C., Li, S., Xin, L., and Yu, P. L. (2024). Choosing the number of factors in factor analysis with incomplete data via a novel hierarchical bayesian information criterion. *Advances in Data Analysis and Classification*, pages 1–27.

Zheng, C. and Chang, H.-H. (2016). High-Efficiency Response Distribution–Based Item Selection Algorithms for Short-Length Cognitive Diagnostic Computerized Adaptive Testing. *Applied Psychological Measurement*, 40(8):608–624.

Zhu, J. and Ma, M. (2012). Uncertainty-Based Active Learning with Instability Estimation for Text Classification. *ACM Transactions on Speech and Language Processing*, 8(4):1–21.

Zhu, N., Wahab, A., Bartušová, M., Ng, N., and Hussain-Alkhateeb, L. (2025). Synthesizing a pragmatic and systemized measure of universal health coverage: verifying the circumstances of mortality categories of death investigated by verbal autopsy. *Frontiers in Public Health*, 13:1422248.

Appendix A

SUPPLEMENT TO CHAPTER 2

A.0.1 Proof of Lemma 1

Let $\boldsymbol{\rho}_{A_j} = (\frac{\theta_i}{\tau_j}, i \in A_j)^T$, $j = 0, \ldots, m$ be the conditional probabilities of each cell within each group A_j and $\boldsymbol{\eta} = (\boldsymbol{\tau}, \boldsymbol{\rho}_{A_j}, j = 0, \ldots, m)$ The probability mass function of $(\boldsymbol{X}, \boldsymbol{Y}) = (X_1, X_2, \ldots, X_k, Y_{A_1}, \ldots, Y_{Am})$ is

$$
\begin{aligned}
f(\boldsymbol{X}, \boldsymbol{Y}|\boldsymbol{\eta}) &= \left(\frac{N!}{\prod_{i=1}^{k} X_i!} \prod_{i=1}^{k} \theta_i^{X_i} \right) \frac{N'!}{\prod_{j=0}^{m} Y_{A_j}!} \prod_{j=0}^{m} \tau_j^{Y_{A_j}} \\
&= \frac{N!}{\prod_{i=1}^{k} X_i!} \frac{N'!}{\prod_{j=0}^{m} Y_{A_j}!} \left\{ \prod_{r \in A_0} (\tau_0 \rho_{A_0}^{(r)})^{X_r} \times \prod_{r \in A_1} (\tau_1 \rho_{A_1}^{(r)})^{X_r} \times \cdots \times \prod_{r \in A_m} (\tau_m \rho_{A_m}^{(r)})^{X_r} \right\} \prod_{j=0}^{m} \tau_j^{Y_{A_j}} \\
&= \frac{N!}{\prod_{i=1}^{k} X_i!} \frac{N'!}{\prod_{j=0}^{m} Y_{A_j}!} \left(\prod_{j=0}^{m} \prod_{r \in A_j} (\tau_j \rho_{A_j}^{(r)})^{X_r} \right) \prod_{j=0}^{m} \tau_j^{Y_{A_j}} \\
&= \frac{N!}{\prod_{i=1}^{k} X_i!} \frac{N'!}{\prod_{j=0}^{m} Y_{A_j}!} \left(\prod_{j=0}^{m} \tau_j^{\sum_{r \in A_j} X_r + Y_{A_j}} \right) \prod_{j=0}^{m} \prod_{r \in A_j} (\rho_{A_j}^{(r)})^{X_r} \\
&= \frac{N!}{\prod_{i=1}^{k} X_i!} \frac{N'}{\prod_{j=0}^{m} Y_{A_j}!} \left(\prod_{j=0}^{m} \tau_j^{X_{A_j} + Y_{A_j}} \right) \prod_{j=0}^{m} \prod_{r \in A_j} (\rho_{A_j}^{(r)})^{X_r}
\end{aligned}
$$

where $X_{A_j} = \sum_{r \in A_j} X_r$.

$$\log\{\prod_{j=0}^{m}\prod_{r\in A_j}(\rho_{A_j}^{(r)})^{X_r}\} = \log\{\prod_{r\in A_0}(\rho_{A_0}^{(r)})^{X_r} \times \prod_{r\in A_1}(\rho_{A_1}^{(r)})^{X_r} \times \cdots \times \prod_{r\in A_m}(\rho_{A_m}^{(r)})^{X_r}\}$$

$$= \sum_{j=0}^{m}\log\prod_{r\in A_j}(\rho_{A_j}^{(r)})^{X_r}$$

$$= \sum_{j=0}^{m}\sum_{r\in A_j}X_r\log(\rho_{A_j}^{(r)})$$

$$= \sum_{j=0}^{m}\sum_{r\in A_j\backslash j^0}X_r\log(\rho_{A_j}^{(r)}) + X_{j^0}\log_{\rho_{A_j}^{j^0}}$$

where $j^0 = A_j^{(0)}$ is the first item of A_j.

For $r \in A_j\backslash j_0$, $j = 0, \ldots, m$:

$$\frac{\partial}{\rho_{A_j}^{(r)}}\log f(\boldsymbol{X}, \boldsymbol{Y}|\boldsymbol{\eta}) = \frac{X_r}{\rho_{A_j}^{(r)}} - \frac{X_{j^0}}{\rho_{A_j}^{(j^0)}}$$

For $r, r' \in A_j\backslash j_0$ and $j = 0, \ldots, m$:

$$\frac{\partial^2}{(\partial\rho_{A_j}^{(r)})(\partial\rho_{A_j}^{(r')})}\log f(\boldsymbol{X}, \boldsymbol{Y}|\boldsymbol{\eta}) = \begin{cases} -\frac{X_r}{(\rho_{A_j}^{(r)})^2} - \frac{X_{j^0}}{(\rho_{A_j}^{j^0})^2} & \text{if } r = r' \\ -\frac{X_r}{(\rho_{A_j}^{(r)})^2} & \text{if } r \neq r' \end{cases}$$

Let $A_p = \{p_1, \ldots, p_{n_p}\}$, $p = 0, \ldots, m$. For the p-th group, the conditional vector of the cells within the group is $\rho_{A_p} = (\frac{\theta_{p_1}}{\tau_p}, \ldots, \frac{\theta_{pn_p}}{\tau_p})$. Considering Jeffrey's prior, we have that.

$$
\begin{aligned}
\pi(\boldsymbol{\eta})d\boldsymbol{\eta} &\propto \sqrt{\left|\mathbb{E}_{\boldsymbol{\theta}}\left[\operatorname{diag}\left(\frac{X_{A_1}+Y_{A_1}}{\tau_1^2},\ldots,\frac{X_{A_m}+Y_{A_m}}{\tau_m^2}\right)+\boldsymbol{i}^{(m)}(\boldsymbol{i}^{(m)})^T\frac{X_{A_0}+Y_{A_0}}{\tau_0^2}\right]\right|} \\
&\times\prod_{j=0}^{m}\sqrt{\left|\mathbb{E}_{\boldsymbol{\theta}}\left[\operatorname{diag}\left(\frac{X_r}{(\rho_{A_j}^{(r)})^2}\right)_{r\in A_j\setminus j_0}+\boldsymbol{i}^{(n_j-1)}(\boldsymbol{i}^{(n_j-1)})^T\frac{X_{j^0}}{(\rho_{A_j}^{(j^0)})^2}\right]\right|} \\
&\propto \sqrt{\left|\operatorname{diag}\left(\frac{1}{\tau_1},\ldots,\frac{1}{\tau_m}\right)+\boldsymbol{i}^{(m)}(\boldsymbol{i}^{(m)})^T\frac{1}{\tau_0}\right|}\times\left(\prod_{j=1}^{m}(\sqrt{\tau_j})^{n_j}\right) \\
&\times\prod_{j=0}^{m}\sqrt{\left|\operatorname{diag}\left(\frac{1}{\rho_{A_j}^{(r)}}\right)_{r\in A_j\setminus j^0}+\boldsymbol{i}^{(n_j-1)}(\boldsymbol{i}^{(n_j-1)})^T\frac{1}{\rho_{A_j}^{(j^0)}}\right|} \\
&=\sqrt{\left|\operatorname{diag}\left(\frac{1}{\tau_1},\ldots,\frac{1}{\tau_m}\right)\right|\left[1+\frac{(\boldsymbol{i}^{(m)})^T}{\sqrt{\tau_0}}\left\{\operatorname{diag}\left(\frac{1}{\tau_1},\ldots,\frac{1}{\tau_m}\right)\right\}^{-1}\frac{\boldsymbol{i}^{(m)}}{\sqrt{\tau_0}}\right]} \\
&\times\left(\prod_{j=1}^{m}\tau_j^{n_j/2}\right)\prod_{j=0}^{m}\sqrt{\left|\operatorname{diag}\left(\frac{1}{\rho_{A_j}^{(r)}}\right)_{r\in A_j\setminus j^0}\right|\left[1+\frac{(\boldsymbol{i}^{(n_j-1)})^T}{\sqrt{\rho_{A_j}^{(j^0)}}}\left\{\operatorname{diag}\left(\frac{1}{\rho_{A_j}^{(r)}}\right)_{r\in A_j\setminus j^0}\right\}^{-1}\frac{\boldsymbol{i}^{(n_j-1)}}{\sqrt{\rho_{A_j}^{(j^0)}}}\right]} \\
&=\left\{\prod_{j=0}^{m}\tau_i^{n_j/2-1}\right\}\prod_{j=0}^{m}\prod_{r\in A_j}(\rho_{A_j}^{(r)})^{1/2-1}
\end{aligned}
$$

The Bayes estimators under the KL loss are the posterior means of $\boldsymbol{\theta}=(\boldsymbol{\tau},\boldsymbol{\rho}_{A_j},j=0,\ldots,m)$. Therefore, concerning the Jeffreys prior and the direct observations $(\boldsymbol{X},\boldsymbol{Y})$, the Bayes estimator is given as

$$
\begin{aligned}
&\hat{\boldsymbol{\theta}}=(\theta_i)_{i=1}^{k},\text{ where for } i\in A_j,\\
&\hat{\theta}_i=\mathbb{E}(\theta_i|\boldsymbol{X},\boldsymbol{Y})=\mathbb{E}(\frac{\theta_i}{\sum_{r\in A_j}\theta_r}|\boldsymbol{X},\boldsymbol{Y})\mathbb{E}(\sum_{r\in A_j}\theta_r|\boldsymbol{X},\boldsymbol{Y})\\
&\quad=\mathbb{E}(\rho_{A_j}^{(i)}|\boldsymbol{X},\boldsymbol{Y})\mathbb{E}(\tau_j|\boldsymbol{X},\boldsymbol{Y})
\end{aligned}
$$

The closed form posterior mean for τ_j and $\rho_{A_j}^{(i)}$ can be obtained as

$$
\mathbb{E}\left(\rho_{A_j}^{(i)}\middle|\boldsymbol{X},\boldsymbol{Y}\right)=\mathbb{E}\left(\frac{\theta_i}{\sum_{r\in A_j}\theta_r}\middle|\boldsymbol{X},\boldsymbol{Y}\right)=\frac{1/2+x_i}{\sum_{r\in A_j}x_r+n_j/2}
$$

$$\mathbb{E}\left(\tau_j|\boldsymbol{X},\boldsymbol{Y}\right) = \mathbb{E}\left(\sum_{r\in A_j}\theta_r\middle|\boldsymbol{X},\boldsymbol{Y}\right) = \frac{y_{A_j}+\sum_{r\in A_j}x_r + n_j/2}{\sum_{l=0}^{m}\{y_{A_l}+\sum_{r\in A_l}(1/2+x_r)\}}$$
$$= \frac{y_{A_j}+\sum_{r\in A_j}x_r + n_j/2}{N+N'+n_j(m+1)/2}$$

Therefore,

$$\hat{\theta}_i = \frac{1/2+x_i}{\sum_{r\in A_j}x_r + n_j/2}\frac{y_{A_j}+\sum_{r\in A_j}x_r + n_j/2}{N+N'+n_j(m+1)/2}$$

We can further show that the posterior of $(\boldsymbol{\eta})$ concerning a prior distribution for $\boldsymbol{\theta}$ being any Dirichlet distribution with parameters $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_k)$ admits a tractable representation.

$$\hat{\theta}_i = \frac{\alpha_i + x_i}{\sum_{r\in A_j}(x_r+\alpha_r)}\frac{y_{A_j}+\sum_{r\in A_j}(\alpha_r+x_r)}{\sum_{l=0}^{m}\{y_{A_l}+\sum_{r\in A_l}(\alpha_r+x_r)\}}$$

A.0.2 Proof of Lemma 2

We would like to compare the risk functions of $\tilde{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\theta}}$.

$$\Delta_{\boldsymbol{\theta}}(N,N') = \mathbb{E}_{\boldsymbol{\theta}}[L(\hat{\boldsymbol{\theta}},\boldsymbol{\theta})] - \mathbb{E}_{\boldsymbol{\theta}}[(L(\tilde{\boldsymbol{\theta}},\boldsymbol{\theta})]$$

For notation simplicity, let $x_{A_j} = \sum_{r\in A_j}(x_r)$, $\theta_{A_j} = \sum_{i\in A_j}\theta_i$, $\alpha_{A_j} = \sum_{r\in A_j}\alpha_r$, and $\alpha_S = \sum_{c=1}^{k}\alpha_c$.

$$
\begin{aligned}
\Delta_{\boldsymbol{\theta}}(N, N') &= \mathbb{E}_{\boldsymbol{\theta}}\left[\sum_{j=0}^{m}\sum_{i\in A_j}\theta_i \log\frac{\tilde{\theta}_i}{\hat{\theta}_i}\right] \\
&= \mathbb{E}_{\boldsymbol{\theta}}\left[\sum_{j=0}^{m}\sum_{i\in A_j}\theta_i \log\left\{\frac{x_i+\alpha_i}{\sum_{c=1}^{k}\alpha_c + x_c}\times\frac{\sum_{r\in A_j}(x_r+\alpha_r)}{\alpha_i+x_i}\frac{\sum_{l=0}^{m}\{y_{A_l}+\sum_{r\in A_l}(\alpha_r+x_r)\}}{y_{A_j}+\sum_{r\in A_j}(\alpha_r+x_r)}\right\}\right] \\
&= \mathbb{E}_{\boldsymbol{\theta}}\left[\sum_{j=0}^{m}\sum_{i\in A_j}\theta_i \log\left\{\frac{N'+N+\sum_{l=0}^{m}\sum_{r\in A_l}(\alpha_r)}{N+\sum_{c=1}^{k}\alpha_c}\times\frac{\sum_{r\in A_j}(x_r+\alpha_r)}{y_{A_j}+\sum_{r\in A_j}(\alpha_r+x_r)}\right\}\right] \\
&= \mathbb{E}_{\boldsymbol{\theta}}\left[\sum_{j=0}^{m}\sum_{i\in A_j}\theta_i \log\left\{\frac{N'+N+\sum_{l=0}^{m}\alpha_{A_l}}{N+\alpha_S}\times\frac{x_{A_j}+\alpha_{A_j}}{y_{A_j}+x_{A_j}+\alpha_{A_j}}\right\}\right] \\
&= \mathbb{E}_{\boldsymbol{\theta}}\left[\log\frac{N'+N+\alpha_S}{N+\alpha_S}+\sum_{j=0}^{m}\theta_{A_j}\log\left\{\frac{x_{A_j}+\alpha_{A_j}}{y_{A_j}+x_{A_j}+\alpha_{A_j}}\right\}\right] \\
&= \sum_{u=1}^{N'}\Delta_{\boldsymbol{\theta}}(N+u-1,1)
\end{aligned}
$$

A.0.3 Proof of Lemma 3

Without loss of generality, we consider $\Delta_{\boldsymbol{\theta}}(N,1)$. Let $(Z_{A_j})_{j=0}^{m} \sim \text{Multi}(1,(\theta_{A_j})_{j=0}^{m})$ be an independent set of multinomial variables.

$$
\begin{aligned}
\Delta_{\boldsymbol{\theta}}(N,1) &= \log\frac{1+N+\alpha_S}{N+\alpha_S}+\mathbb{E}_{\boldsymbol{\theta}}\left[\sum_{j=0}^{m}\theta_{A_j}\log\left\{\frac{x_{A_j}+\alpha_{A_j}}{z_{Aj}+x_{A_j}+\alpha_{A_j}}\right\}\right] \\
&= \log\frac{1+N+\alpha_S}{N+\alpha_S}+\mathbb{E}_{\boldsymbol{\theta}}\left[\sum_{j=0}^{m}\theta_{A_j}\log\left\{\frac{x_{A_j}+\alpha_{A_j}}{z_{A_j}+x_{A_j}+\alpha_{A_j}}\right\}\middle| Z_{A_j}=1\right]\mathbb{P}(Z_{A_j}=1) \\
&+ \mathbb{E}_{\boldsymbol{\theta}}\left[\sum_{j=0}^{m}\theta_{A_j}\log\left\{\frac{x_{A_j}+\alpha_{A_j}}{z_{A_j}+x_{A_j}+\alpha_{A_j}}\right\}\middle| Z_{A_j}=0\right]\mathbb{P}(Z_{A_j}=0) \\
&= \log\frac{1+N+\alpha_S}{N+\alpha_S}+\mathbb{E}_{\boldsymbol{\theta}}\left[\sum_{j=0}^{m}\theta_{A_j}^2\log\left\{\frac{x_{A_j}+\alpha_{A_j}}{1+x_{A_j}+\alpha_{A_j}}\right\}\right] \\
&= \log\left\{1+\frac{1}{N+\alpha_S}\right\}+\mathbb{E}_{\boldsymbol{\theta}}\left[\sum_{j=0}^{m}\theta_{A_j}^2\log\left\{1-\frac{1}{1+x_{A_j}+\alpha_{A_j}}\right\}\right]
\end{aligned}
$$

where the second equality follows because $Z_{A_j} \sim \text{Bern}(\theta_{A_j})$, $\mathbb{P}(Z_{A_j}=1)=\theta_{A_j}$

Define the function $G : (0,1) \to (0,\infty)$ by:

$$G(\theta) = \theta^2 \mathbb{E}\left[-\log\left\{1 - \frac{1}{1 + X(\theta) + \alpha_{A_j}}\right\}\right],\ \theta \in (0,1)$$

where $X(\theta) \sim \text{Bin}(N,\theta)$ for $\theta \in (0,1)$.

Then we have

$$\Delta_{\boldsymbol{\theta}}(N,1) = \log\left\{1 + \frac{1}{N + \alpha_S}\right\} - \sum_{j=0}^{m} G(\theta_{A_j})$$

If we can prove the condition required for G to be convex, then the maximum is attained.

We first prove the following results:

1. Suppose $X \sim \text{Binom}(n,\theta)$, and let $h(X)$ be a function of X. Then we can show that

$$\frac{\partial}{\partial\theta}\mathbb{E}[h(X)] = \frac{1}{\theta}\mathbb{E}[X\{h(X) - h(X-1)\}]$$

Proof: Using the definition of the expectation of a binomial random variable, and the lemma of Johnson (1987), we have that

$$\begin{aligned}\frac{\partial}{\partial\theta}\mathbb{E}[h(X)] &= \mathbb{E}\left\{\left(\frac{X}{\theta} - \frac{n-X}{1-\theta}\right)h(X)\right\} \\ &= \frac{1}{\theta(1-\theta)}\mathbb{E}\left\{(X - n\theta)\,h(X)\right\} \\ &= \frac{1}{\theta}\mathbb{E}\left\{X(h(X) - h(X-1))\right\}\end{aligned}$$

2. Suppose $X \sim \text{Binom}(n,\theta)$, and let $h(X)$ be a function of X. Then we can show that

$$\frac{\partial}{\partial\theta^2}\{\theta^2\mathbb{E}[h(X)]\} = \frac{1}{\theta}\mathbb{E}\left[X\left\{(X+1)h(X) - 2Xh(X-1) + (X-1)h(X-2)\right\}\right]$$

Proof:

$$\begin{aligned}\frac{\partial^2}{\partial\theta^2}\{\theta\mathbb{E}[h(X)]\} &= \frac{\partial}{\partial\theta}[\mathbb{E}[h(X)] + \mathbb{E}[X\{h(X) - h(X-1)\}] \\ &= \frac{\partial}{\partial\theta}\mathbb{E}[(X+1)h(X) - Xh(X-1)] \\ &= \frac{1}{\theta}\mathbb{E}[X\{(X+1)h(X) - 2Xh(X-1) + (X-1)h(X-2)\}]\end{aligned}$$

Consider the function $g : [0, \infty) \to \mathbb{R}$ and let $g(x) = 0$ for $x \in (-\infty, 0)$. Let $X = X(\theta)$ for simplicity of the notation. Using the results from the foregoing lemmas, we have

$$\begin{aligned}
\frac{\partial^2}{\partial\theta^2}\mathbb{E}[g(X)] &= \frac{\partial}{\partial\theta}\left(2\theta\mathbb{E}[g(X)] + \theta\mathbb{E}[X\{g(X) - g(X-1)\}]\right) \\
&= 2\mathbb{E}[g(X)] + 2\mathbb{E}[X\{g(X) - g(X-1)\}] + \mathbb{E}[X\{g(X) - g(X-1)\}] \\
&+ \theta\frac{\partial}{\partial\theta}\mathbb{E}[X\{g(X) - g(X-1)\}] \\
&= 2\mathbb{E}[g(X)] + 2\mathbb{E}[X\{g(X) - g(X-1)\}] + \mathbb{E}[X\{g(X) - g(X-1)\}] \\
&+ \mathbb{E}\left[X\{X(g(X) - g(X-1)) - (X-1)(g(X-1) - g(X-2))\}\right] \\
&= \mathbb{E}[(X^2 + 3X + 2)g(X) - (2X^2 + 2X)g(X-1) + (X^2 - X)g(X-2)] \\
&= \mathbb{E}[\tilde{g}(X) - 2\tilde{g}(X-1) + \tilde{g}(X-2)]
\end{aligned}$$

where $\tilde{g}(x) = (x+2)(x+1)g(x)$ for $x \in \mathbb{R}$.

Returning to the definition of $G(\theta)$, since $\alpha_{A_j} = \sum_{i \in A_j} \alpha_i$ which only depend on $|A_j|$ but not θ_{A_j}, we write $\tilde{\alpha} = \alpha_{A_j}$ for notation simplicity. Suppose that

$$g(x) = -\log\left\{1 - \frac{1}{1 + x + \tilde{\alpha}}\right\}$$

for all $x \in [0, \infty)$.

Fixing $x \in [0, \infty)$, we have that

$$\tilde{g}(x) = (x^2 + 3x + 2)\log(\frac{1 + x + \tilde{\alpha}}{x + \tilde{\alpha}})$$

$$\tilde{g}'(x) = (2x+3)\log\frac{1 + x + \tilde{\alpha}}{x + \tilde{\alpha}} - \frac{(x^2 + 3x + 2)}{(1 + x + \tilde{\alpha})(x + \tilde{\alpha})}$$

$$\tilde{g}''(x) = 2\log\frac{1 + x + \tilde{\alpha}}{x + \tilde{\alpha}} - 2\times\frac{(2x+3)}{(1 + x + \tilde{\alpha})(x + \tilde{\alpha})} + \frac{x^2 + 3x + 2}{(1 + x + \tilde{\alpha})^2(x + \tilde{\alpha})} + \frac{(x^2 + 3x + 2)}{(1 + x + \tilde{\alpha})(x + \tilde{\alpha})^2}$$

Hence by the low-integer-order polylogarithms series property $\sum_{k=1}^{\infty} \frac{1}{k} z^k = -\log(1-z)$ we have that

$$
\begin{aligned}
(1+x+\tilde{\alpha})\tilde{g}''(x) &= 2\sum_{k=1}^{\infty}\frac{1}{k}\frac{1}{(1+x+\tilde{\alpha})^{k-1}} - 2\times\frac{2x+3}{x+\tilde{\alpha}} + \frac{x^2+3x+2}{(1+x+\tilde{\alpha})(x+\tilde{\alpha})} + \frac{(x^2+3x+2)}{(x+\tilde{\alpha})^2} \\
&= 2 + 2\sum_{k=2}^{\infty}\frac{1}{k}\frac{1}{(1+x+\tilde{\alpha})^{k-1}} - 4\times(1-\frac{\tilde{\alpha}-3/2}{x+\tilde{\alpha}}) \\
&+ \left(1-\frac{\tilde{\alpha}-1}{1+x+\tilde{\alpha}}+1-\frac{\tilde{\alpha}-2}{x+\tilde{\alpha}}\right)\left(1-\frac{\tilde{\alpha}-1}{x+\tilde{\alpha}}\right) \\
&= 2\sum_{k=2}^{\infty}\frac{1}{k}\frac{1}{(1+x+\tilde{\alpha})^{k-1}} + 4\times\frac{\tilde{\alpha}-3/2}{x+\tilde{\alpha}} \\
&- \left(\frac{\tilde{\alpha}-1}{1+x+\tilde{\alpha}}+\frac{\tilde{\alpha}-2}{x+\tilde{\alpha}}\right) - 2\times\frac{\tilde{\alpha}-1}{x+\tilde{\alpha}} + \left(\frac{\tilde{\alpha}-1}{1+x+\tilde{\alpha}}+\frac{\tilde{\alpha}-2}{x+\tilde{\alpha}}\right)\frac{\tilde{\alpha}-1}{x+\tilde{\alpha}} \\
&= 2\sum_{k=2}^{\infty}\frac{1}{k}\frac{1}{(x+1+\tilde{\alpha})^{k-1}} - \frac{1}{x+\tilde{\alpha}} + \frac{\tilde{\alpha}-1}{(x+\tilde{\alpha})(1+x+\tilde{\alpha})} + \left(\frac{\tilde{\alpha}-1}{1+x+\tilde{\alpha}}+\frac{\tilde{\alpha}-2}{x+\tilde{\alpha}}\right)
\end{aligned}
$$

If $\tilde{\alpha} \geq 2$, then

$$
\begin{aligned}
(1+x+\tilde{\alpha})\tilde{g}''(x) &\geq 2\times\frac{1}{2}\times\frac{1}{1+x+\tilde{\alpha}} - \frac{1}{x+\tilde{\alpha}} + \frac{\tilde{\alpha}-1}{(x+\tilde{\alpha})(1+x+\tilde{\alpha})} \\
&= -\frac{1}{(x+\tilde{\alpha})(1+x+\tilde{\alpha})} + \frac{\tilde{\alpha}-1}{(x+\tilde{\alpha})(1+x+\tilde{\alpha})} \geq 0
\end{aligned}
$$

Therefore, $\tilde{g}(X) - 2\tilde{g}(X-1) + \tilde{g}(X-2) \geq 0$ when $X \geq 2$.

For $X = 1$, we have

$$
\begin{aligned}
\frac{1}{6}(\tilde{g}(1) - 2\tilde{g}(0) + \tilde{g}(-1)) &= \log\left(1+\frac{1}{1+\tilde{\alpha}}\right) - \frac{2}{3}\log\left(1+\frac{1}{\tilde{\alpha}}\right) \\
&\geq \log\left(1+\frac{1}{1+\tilde{\alpha}}\right) - \log\left(1+\frac{2/3}{\tilde{\alpha}}\right) \geq 0
\end{aligned}
$$

For $X = 0$, we have $\tilde{g}(0) - 2\tilde{g}(-1) + \tilde{g}(-2) \geq 0$ trivially. Then we have shown that

$$
\frac{\partial^2}{\partial\theta^2}\theta^2\mathbb{E}[g(X)] = \mathbb{E}[\tilde{g}(X) - 2\tilde{g}(X-1) + \tilde{g}(X-2)] \geq 0
$$

Hence when $\alpha_{A_j} \geq 2$, the function G is convex.

Assuming $\alpha_{A_j} \geq 2$, because G is convex and $\Delta_{\boldsymbol{\theta}}(N,1)$ has a convex domain, the maximum value exists and can be attained. We would like to find the $\boldsymbol{\theta}^*$ such that the risk difference $\Delta_{\boldsymbol{\theta}}(N,1)$ is maximized.

Let $\boldsymbol{\theta}^* = (\theta_i^*)_{i=0}^k \in \boldsymbol{\Theta}$, where for $i \in A_j = \{j_1, \ldots, j_{n_j}\}$, $\theta^*_{i:i\in A_j} = \frac{1}{(1+m)n_j}$ for all $j = 0, \ldots, m$. Then $\sum_{i=0}^k \theta^*_{i:i\in A_j} = \theta_{A_j} = \frac{1}{m+1}$ for all $j = 0, 1, \ldots, m$.

By Jensen's inequality for a real convex function, we have that

$$\phi(\frac{\sum a_i x_i}{\sum a_i}) \leq \frac{\sum a_i \phi(x_i)}{\sum a_i}$$

And as a particular case, if a_i are equal, we have

$$\phi(\frac{\sum x_i}{n}) \leq \frac{\sum \phi(x_i)}{n}$$

Therefore,

$$\begin{aligned}
\Delta_{\boldsymbol{\theta}}(N,1) &= \log\left\{1 + \frac{1}{N+\alpha_S}\right\} - \sum_{j=0}^{m} G(\theta_{A_j}) \\
&= \log\left\{1 + \frac{1}{N+\alpha_S}\right\} - (1+m)\frac{1}{1+m}\sum_{j=0}^{m} G(\theta_{A_j}) \\
&\leq \log\left\{1 + \frac{1}{N+\alpha_S}\right\} - (1+m)G(\frac{1}{m+1}\sum_{j=1}^{m} \theta_{A_j}) \\
&= \log\left\{1 + \frac{1}{N+\alpha_S}\right\} - \sum_{i=0}^{m} G(\frac{1}{m+1}) \\
&= \Delta_{\boldsymbol{\theta}^*}(N,1)
\end{aligned}$$

With the foregoing results, we attempt to obtain the condition for the dominance between $\hat{\boldsymbol{\theta}}$ and $\tilde{\boldsymbol{\theta}}$:

$$\begin{aligned}
\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\Delta_{\boldsymbol{\theta}}(N,N') &= \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\sum_{u=1}^{N'}\Delta_{\boldsymbol{\theta}}(N+u-1,1) \\
&\leq \sum_{u=1}^{N'}\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\Delta_{\boldsymbol{\theta}}(N+u-1,1) \\
&= \sum_{u=1}^{N'}\Delta_{\boldsymbol{\theta}^*}(N+u-1,1) \\
&= \Delta_{\boldsymbol{\theta}^*}(N,N')
\end{aligned}$$

Then

$$\begin{aligned}
\Delta_{\boldsymbol{\theta}^*}(N,N') &= \mathbb{E}_{\boldsymbol{\theta}^*}\left[\log\frac{N'+N+\alpha_S}{N+\alpha_S}+\sum_{j=0}^{m}\frac{1}{m+1}\log\left\{\frac{x_{A_j}+\alpha_{A_j}}{y_{A_j}+x_{A_j}+\alpha_{A_j}}\right\}\right] \\
&= -\mathbb{E}_{\boldsymbol{\theta}^*}\left[\sum_{j=0}^{m}\frac{1}{m+1}\log\left\{\frac{y_{A_j}+x_{A_j}+\alpha_{A_j}}{x_{A_j}+\alpha_{A_j}}\right\}+\log\frac{N+\alpha_S}{N'+N+\alpha_S}\right] \\
&= -\mathbb{E}_{\boldsymbol{\theta}^*}\left[\sum_{j=0}^{m}\frac{1}{m+1}\log\left\{\frac{y_{A_j}+x_{A_j}+\alpha_{A_j}}{x_{A_j}+\alpha_{A_j}}\times\frac{N+\alpha_S}{N+N'+\alpha_S}\right\}\right] \\
&= -\mathbb{E}_{\boldsymbol{\theta}^*}\left[\sum_{j=0}^{m}\frac{1}{m+1}\log\left\{\frac{y_{A_j}/N'+(x_{A_j}+\alpha_{A_j})/N'}{1+(N+\alpha_S)/N'}\Big/\frac{x_{A_j}+\alpha_{A_j}}{N+\alpha_S}\right\}\right] \\
&\to -\mathbb{E}_{\boldsymbol{\theta}^*}\left[\sum_{j=0}^{m}\frac{1}{m+1}\log\left\{\frac{1}{m+1}\Big/\frac{x_{A_j}+\alpha_{A_j}}{N+\alpha_S}\right\}\right]
\end{aligned}$$

as $N'\to\infty$ by law of large numbers. We can show that the right-hand side is negative because

$$\mathbb{P}_{\theta^*}\left\{\left(\frac{x_{A_j}+\alpha_{A_j}}{N+\alpha_S}\right)_{j=0}^{m}\neq\left(\frac{1}{m+1}\right)_{j=0}^{m}\right\}>0$$

Therefore,

$$\Delta_{\boldsymbol{\theta}^*}(N,N')\leq\sup\Delta_{\boldsymbol{\theta}}(N,N')<0$$

for sufficiently large N'.

A.0.4 Proof of Theorem 3

Assume $N' = 1$ without loss of generality. Then it is sufficient to show that

$$\Delta_{\boldsymbol{\theta}^*}(N, 1) = \log\left\{1 + \frac{1}{N + \alpha_S}\right\} - (m+1)G(\frac{1}{m+1})$$

is negative when N is sufficiently large.

Let $\theta^*_{A_j} = \frac{1}{1+m}$ and $X \sim \text{Bin}(N, \theta^*_{A_j})$

$$\begin{aligned}(1+m)G(\frac{1}{1+m}) &= \frac{1}{1+m}\mathbb{E}\left[-\log\left\{1 - \frac{1}{1+X+\alpha_{A_j}}\right\}\right] \\ &= \theta^*_{A_j}\sum_{k=1}^{\infty}\frac{1}{k}\mathbb{E}\left[\frac{1}{(1+X+\alpha_{A_j})^k}\right]\end{aligned}$$

By (3) of Cressie et al. (1981), which is

$$\mathbb{E}(X^{-n}) = \Gamma(n)^{-1}\int_0^{\infty} t^{n-1}M_X(-t)dt$$

We have that

$$\begin{aligned}\frac{1}{k}\mathbb{E}\left[\frac{1}{(1+X+\alpha_{A_j})^k}\right] &= \frac{1}{k}\frac{1}{\Gamma(k)}\int_0^{\infty} t^{k-1}\mathbb{E}[e^{-(1+X+\alpha_{A_j})t}]dt \\ &= \frac{1}{k!}\int_0^{\infty} t^{k-1}e^{-(1+\alpha_{A_j})t}\mathbb{E}[e^{-Xt}]dt\end{aligned}$$

for $k = 1, 2, 3, \ldots$. Then

$$\begin{aligned}(m+1)G(\frac{1}{m+1}) &= \theta^*_{A_j}\int_0^{\infty}\sum_{k=1}^{\infty}\frac{t^{k-1}}{k!}e^{-(1+\alpha_{A_j})t}\mathbb{E}[e^{-Xt}]dt \\ &= \theta^*_{A_j}\int_0^{\infty}\frac{e^t - 1}{t}e^{-(1+\alpha_{A_j})t}\mathbb{E}[e^{-Xt}]dt\end{aligned}$$

Note that $\mathbb{E}[e^{-Xt}] = (1 - \theta^*_{A_j} + \theta^*_{A_j}e^{-t})^N$ for all $t \in (0, \infty)$ by the moment generating function of binomial distribution. Then we have

$$(m+1)G(\frac{1}{m+1}) = \theta^*_{A_j}\int_0^\infty \frac{e^t-1}{t}e^{-(1+\alpha_{A_j})t}(1-\theta^*_{A_j}+\theta^*_{A_j}e^{-t})^N dt$$

Also,

$$\begin{aligned}\log\left\{1+\frac{1}{N+\alpha_S}\right\} = -\log\left\{1-\frac{1}{1+N+\alpha_S}\right\} &= \sum_{k=1}^\infty \frac{1}{k}\frac{1}{(N+1+\alpha_S)^k}\\ &= \sum_{k=1}^\infty \frac{1}{k!}\int_0^\infty t^{k-1}e^{-(N+1+\alpha_S)t}dt\\ &= \int_0^\infty \frac{e^t-1}{t}e^{-(N+1)t}e^{-\alpha_S t}\end{aligned}$$

Therefore,

$$\Delta_{\boldsymbol{\theta}^*}(N,1) = \int_0^\infty \frac{e^t-1}{t}e^{-(N+1)t}e^{-\alpha_S t} - \theta^*_{A_j}\int_0^\infty \frac{e^t-1}{t}e^{-(1+\alpha_{A_j})t}(1-\theta^*_{A_j}+\theta^*_{A_j}e^{-t})^N dt$$

Note that

$$\begin{aligned}0 \le \left(\frac{e^t-1}{t}\right) &= \sum_{k=1}^\infty \frac{t^{k-1}}{k!} \le \sum_{k=1}^\infty \frac{t^{k-1}}{(k-1)!} = e^t\\ 0 \le \left(\frac{e^t-1}{t}\right)' &= \sum_{k=2}^\infty \frac{(k-1)t^{k-2}}{k!} \le \sum_{k=2}^\infty \frac{t^{k-2}}{(k-2)!} = e^t\\ 0 \le \left(\frac{e^t-1}{t}\right)'' &= \sum_{k=3}^\infty \frac{(k-1)(k-2)t^{k-3}}{k!} \le \sum_{k=3}^\infty \frac{t^{k-3}}{(k-3)!} = e^t\end{aligned}$$

and

$$\begin{aligned}\left(\frac{e^t-1}{t}\right) &\le \frac{e^t}{t}\\ \left(\frac{e^t-1}{t}\right)' &= \frac{e^t}{t} - \frac{e^t-1}{t^2} \le \frac{e^t}{t}\\ \left(\frac{e^t-1}{t}\right)'' &= \frac{e^t}{t} - 2\frac{e^t}{t^2} + 2\frac{e^t-1}{t^3} \le \frac{e^t}{t} + 2\frac{e^t}{t^3}\end{aligned}$$

By integration by parts, we have

$$
\begin{aligned}
(N+1)\Delta_{\boldsymbol{\theta}^*}(N,1) &= \left[-\frac{e^t-1}{t}e^{-(N+1)t}e^{-\alpha_S t}\right]_0^\infty - \left[-\frac{e^t-1}{t}e^{-\alpha_{A_j}t}(1+\theta^*_{A_j}+\theta^*_{A_j}e^{-t})^{N+1}\right]_0^\infty \\
&+ \int_0^\infty \left\{\left(\frac{e^t-1}{t}\right)' - \left(\frac{e^t-1}{t}\right)\alpha_S\right\} e^{-(N+1)t}e^{-\alpha_S t}dt \\
&- \int_0^\infty \left\{\left(\frac{e^t-1}{t}\right)' - \left(\frac{e^t-1}{t}\right)\alpha_{A_j}\right\} e^{-\alpha_{A_j}t}(1-\theta^*_{A_j}+\theta^*_{A_j}e^{-t})^{N+1}dt \\
&= \int_0^\infty h_1(t)e^{-(N+1)t}e^{-\alpha_S t}dt - \int_0^\infty h_2(t)e^{-\alpha_{A_j}t}(1-\theta^*_{A_j}+\theta^*_{A_j}e^{-t})^{N+1}dt
\end{aligned}
$$

where $h_1(t) = \left(\frac{e^t-1}{t}\right)' - \left(\frac{e^t-1}{t}\right)\alpha_S$ and $h_2(t) = \left(\frac{e^t-1}{t}\right)' - \left(\frac{e^t-1}{t}\right)\alpha_{A_j}$ for $t \in (0,\infty)$.

By integration by parts, we have

$$
\begin{aligned}
(N+1)^2\Delta_{\boldsymbol{\theta}^*}(N,1) &= \left[h_1(t)e^{-(N+1)t}e^{-\alpha_S t}\right]_0^\infty \\
&- \frac{N+1}{N+2}\left[\frac{1}{\theta^*_{A_j}}h_2(t)e^{-(\alpha_{A_j}-1)t}(1-\theta^*_{A_j}+\theta^*_{A_j}e^{-t})^{N+2}\right]_0^\infty \\
&+ \int_0^\infty \left\{h_1'(t) - h_1(t)\alpha_S\right\} e^{-(N+1)t}e^{-\alpha_S t}dt \\
&- \frac{N+1}{N+2}\int_0^\infty \frac{1}{\theta^*_{A_j}}\left[h_2'(t) - h_2(t)(\alpha_{A_j}-1)\right] e^{-(\alpha_{A_j}-1)t}(1-\theta^*_{A_j}+\theta^*_{A_j}e^{-t})^{N+2}dt \\
&= h_1(0) - \frac{N+1}{N+2}\frac{1}{\theta^*_{A_j}}h_2(0) + \int_0^\infty \left[h_1'(t) - h_1(t)\alpha_S\right] e^{-(N+1)t}e^{-\alpha_S t}dt \\
&- \frac{N+1}{N+2}\int_0^\infty \frac{1}{\theta^*_{A_j}}\left[h_2'(t) - h_2(t)(\alpha_{A_j}-1)\right] e^{-(\alpha_{A_j}-1)t}(1-\theta^*_{A_j}+\theta^*_{A_j}e^{-t})^{N+2}dt
\end{aligned}
$$

Consider the following

$$
\begin{aligned}
\left|\left[h_2'(t) - h_2(t)(\alpha_{A_j}-1)\right] e^{-(\alpha_{A_j}-1)t}\right| &\le \left[e^t + e^t\alpha_{A_j}(\alpha_{A_j}-1)\right] e^{-(\alpha_{A_j}-1)t} \\
&= [1+\alpha_{A_j}(\alpha_{A_j}-1)]e^{-(\alpha_{A_j}-2)t}
\end{aligned}
$$

By assumption we have $\alpha_{A_j} \ge 2$.

For $\alpha_{A_j} > 2$, when $t \in (0,1]$, we have

$$\left| \left[h_2'(t) - h_2(t)(\alpha_{A_j} - 1)\right] e^{-(\alpha_{A_j}-1)t} \right| \leq \left| h_2'(t) - h_2(t) \right| = 6e^t$$

and when $t \in (1, \infty)$ we have

$$\left| \left[h_2'(t) - h_2(t)(\alpha_{A_j} - 1)\right] e^{-(\alpha_{A_j}-1)t} \right| \leq \frac{3}{t^2} + \frac{7}{t} e^{-t}$$

For $\alpha_{A_j} = 2$, by the dominated convergence theorem, as $N \to \infty$,

$$(N+1)^2 \Delta_{\boldsymbol{\theta}^*}(N, 1) \to h_1(0) - \frac{1}{\theta^*_{A_j}} h_2(0) = \frac{1}{2}(1 - \frac{1}{\theta^*_{A_j}}) < 0$$

The maximum risk difference is

$$\begin{aligned} \Delta_{\boldsymbol{\theta}^*}(N, 1) &= \log\left\{1 + \frac{1}{N + \alpha_S}\right\} - (m+1) G(\frac{1}{m+1}) \\ &= \log\left\{1 + \frac{1}{N + \alpha_S}\right\} - \frac{1}{m+1} \mathbb{E}\left[-\log\left\{1 - \frac{1}{1 + X + \alpha_{A_j}}\right\}\right] \end{aligned}$$

A.0.5 Additional Results

Additional simulation scenario: $S = \{1, 2, \ldots, 9\}$, $A_0 = \{1, 2, 3, 4, 5\}$, $A_1 = \{5,6, 7, 8, 9\}$

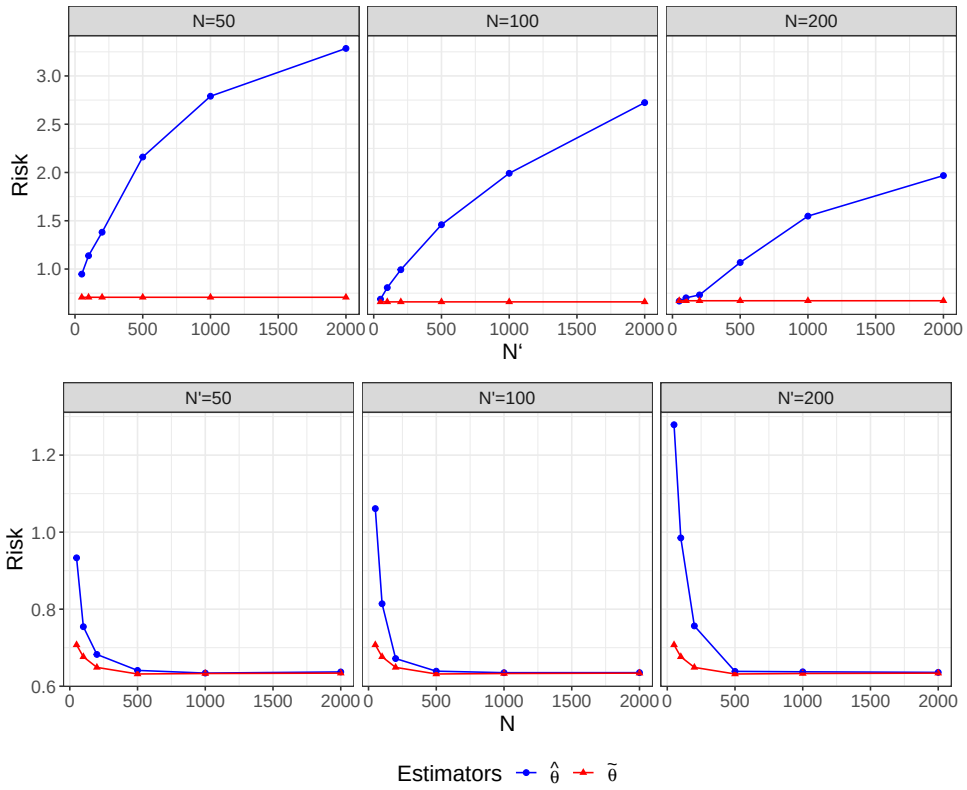


Figure A.1: Comparison of risk between $\hat{\boldsymbol{\theta}}_2$ and $\tilde{\boldsymbol{\theta}}_2$ estimators in an additional simulation scenario in which partial categorizations are overlapping unions of the complete categories. We report the Bayes risk fixing the completely classified sample size N in each subplot while increasing the partially classified sample size N' in the first row, and fixing N' while increasing N in the second row.

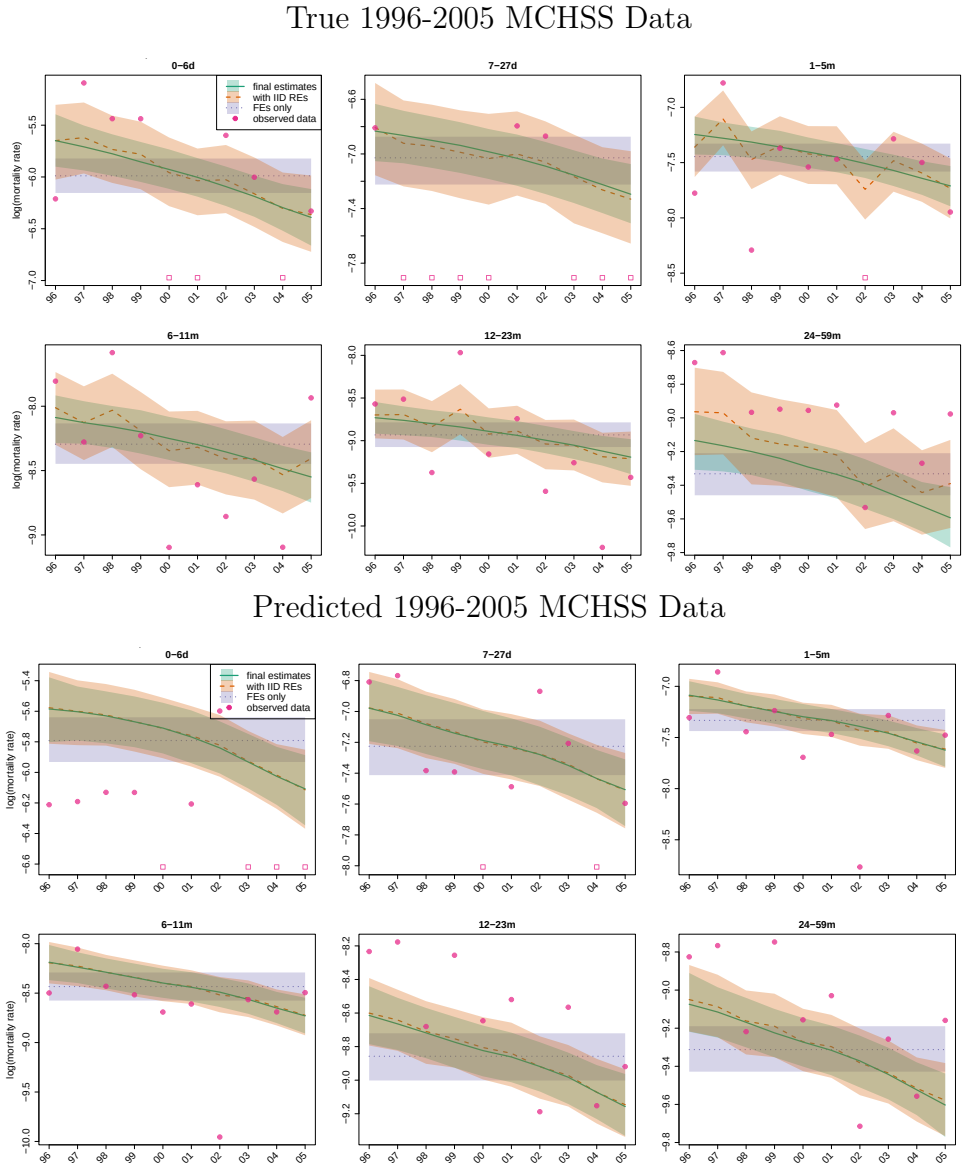


Figure A.2: Selected results from the MCHSS data showing empirical data, estimated posterior medians, and posterior 80% intervals for log mortality rates of other non-communicable diseases in the east urban region. Combinations with no deaths are represented by an open square.

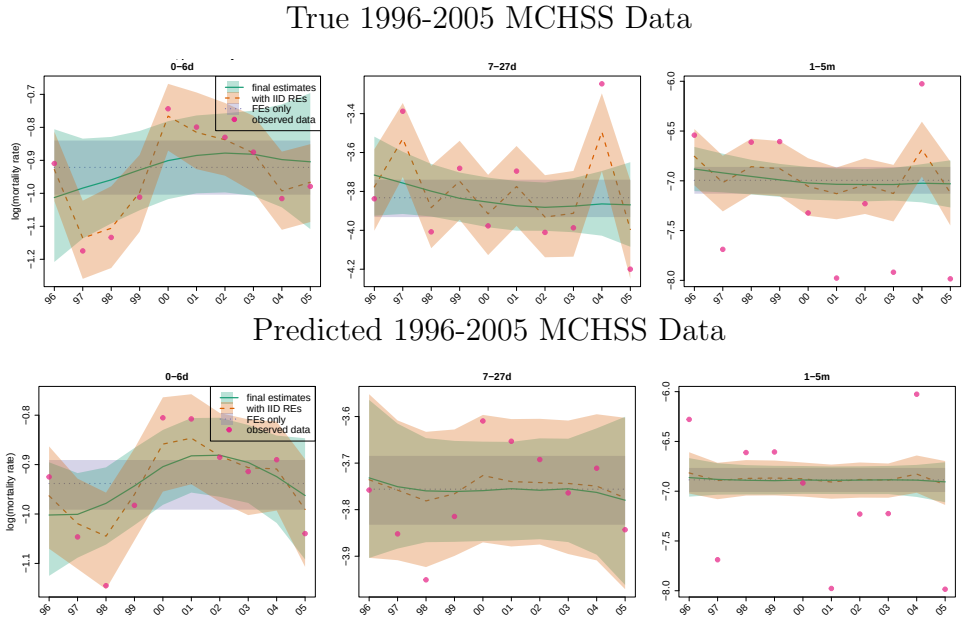


Figure A.3: Selected results from the MCHSS data showing empirical data, estimated posterior medians, and posterior 80% intervals for log mortality rates of prematurity in the west rural region. Combinations with no deaths are represented by an open square.

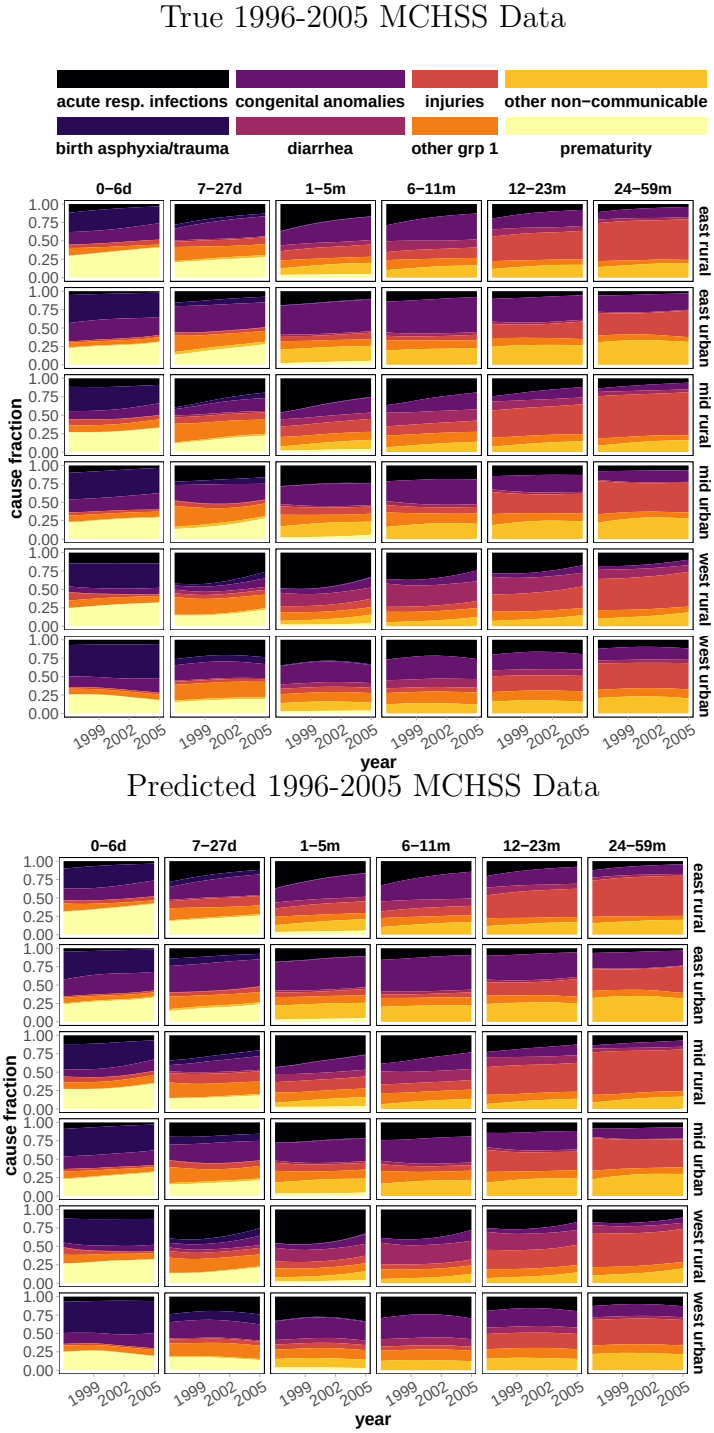


Figure A.4: Comparisons of estimated CSMFs between models based on the true MCHSS data and the estimated MCHSS data using the proposed data reconciliation method.

Appendix B

SUPPLEMENT TO CHAPTER 3

B.1 Additional results for the synthetic data examples

Tables B.1 and B.2 contain more extensive results on the classification accuracy and questionnaire length for the synthetic data under the correctly specified model and misspecified models, respectively.

Table B.1: Classification accuracy and questionnaire length for the synthetic data under the correctly specified model. Median and 5th and 95th percentiles of the questionnaire length are shown for each stopping rule conditions.

p_{1st}	d	Stopping Rule	Acc	Median	Lower	Upper
0.8	0.75	Point Est	0.84	5	3	14
		Pred $r = 0.5$	0.92	10	4	50
		Pred $r = 0.7$	0.95	13	6	50
0.8	0.5	Point Est	0.85	5	3	14
		Pred $r = 0.5$	0.92	10	4	50
		Pred $r = 0.7$	0.95	14	6	50
0.8	0.25	Point Est	0.88	6	3	16
		Pred $r = 0.5$	0.94	10	4	50
		Pred $r = 0.7$	0.96	15	6	50
0.8	0	Point Est	0.96	8	5	35
		Pred $r = 0.5$	0.96	13	6	50
		Pred $r = 0.7$	0.96	20	7	50
0.9	0.75	Point Est	0.95	7	5	17

		Pred $r = 0.5$	0.95	12	6	50
		Pred $r = 0.7$	0.96	17	7	50
0.9	0.5	Point Est	0.95	7	5	18
		Pred $r = 0.5$	0.95	12	6	50
		Pred $r = 0.7$	0.96	18	7	50
0.9	0.25	Point Est	0.95	7	5	24
		Pred $r = 0.5$	0.96	12	6	50
		Pred $r = 0.7$	0.96	18	7	50
0.9	0	Point Est	0.97	10	7	50
		Pred $r = 0.5$	0.96	16	8	50
		Pred $r = 0.7$	0.96	23	8	50
0.95	0.75	Point Est	0.97	9	6	24
		Pred $r = 0.5$	0.96	13	7	50
		Pred $r = 0.7$	0.96	18	8	50
0.95	0.5	Point Est	0.97	9	6	35
		Pred $r = 0.5$	0.96	13	7	50
		Pred $r = 0.7$	0.96	20	8	50
0.95	0.25	Point Est	0.97	10	6	50
		Pred $r = 0.5$	0.96	15	7	50
		Pred $r = 0.7$	0.97	22	8	50
0.95	0	Point Est	0.98	12	8	50
		Pred $r = 0.5$	0.96	20	8	50
		Pred $r = 0.7$	0.97	26	9	50

Table B.2: Classification accuracy and questionnaire length for the synthetic data under the misspecified model. Median and 5th and 95th percentiles of the questionnaire length are shown for each stopping rule conditions.

p_{1st}	d	Stopping Rule	Acc	Median	Lower	Upper
0.8	0.75	Point Est	0.86	4	3	11
		Pred $r = 0.5$	0.88	4	3	18
		Pred $r = 0.7$	0.94	5	3	25
0.8	0.5	Point Est	0.88	5	3	11
		Pred $r = 0.5$	0.88	4	3	18
		Pred $r = 0.7$	0.94	5	3	27
0.8	0.25	Point Est	0.90	5	3	11
		Pred $r = 0.5$	0.90	4	3	22
		Pred $r = 0.7$	0.94	6	3	30
0.8	0	Point Est	0.96	6	5	17
		Pred $r = 0.5$	0.94	6	4	30
		Pred $r = 0.7$	0.99	8	5	50
0.9	0.75	Point Est	0.94	6	4	12
		Pred $r = 0.5$	0.96	6	4	23
		Pred $r = 0.7$	0.98	7	4	50
0.9	0.5	Point Est	0.94	6	4	13
		Pred $r = 0.5$	0.96	6	4	23
		Pred $r = 0.7$	0.98	7	4	50
0.9	0.25	Point Est	0.94	6	4	16
		Pred $r = 0.5$	0.96	6	4	23
		Pred $r = 0.7$	0.99	7	4	50
0.9	0	Point Est	0.96	7	6	20
		Pred $r = 0.5$	0.98	8	5	42

		Pred $r = 0.7$	0.99	9	5	50
0.95	0.75	Point Est	0.96	7	5	16
		Pred $r = 0.5$	0.98	7	5	33
		Pred $r = 0.7$	0.98	8	5	50
0.95	0.5	Point Est	0.96	7	5	17
		Pred $r = 0.5$	0.98	7	5	33
		Pred $r = 0.7$	0.99	9	5	50
0.95	0.25	Point Est	0.96	7	5	17
		Pred $r = 0.5$	0.98	8	5	34
		Pred $r = 0.7$	0.99	9	5	50
0.95	0	Point Est	0.98	8	6	25
		Pred $r = 0.5$	0.99	8	6	50
		Pred $r = 0.7$	0.99	10	6	50

B.2 Additional results for the PHMRC data example

B.2.1 Descriptive statistics of the PHMRC dataset

Table B.3 summarizes the sample size of each COD in the PHMRC data.

Table B.3: Sample size of each COD in the PHMRC dataset.

COD	Sample Size
Stroke	630
Other Non-communicable Diseases (NCD)	599
Pneumonia	540
AIDS	502
Maternal	468

Other Cardiovascular Diseases	416
Renal Failure	416
Diabetes	414
Acute Myocardial Infarction	400
Cirrhosis	313
TB	276
Other Infectious Diseases	263
Diarrhea/Dysentery	228
Road Traffic	202
Breast Cancer	195
Falls	173
COPD	171
Homicide	167
Leukemia/Lymphomas	156
Cervical Cancer	155
Suicide	124
Fires	122
Drowning	106
Lung Cancer	106
Other Injuries	103
Malaria	100
Colorectal Cancer	99
Poisonings	86
Bite of Venomous Animal	66
Stomach Cancer	62
Epilepsy	48
Prostate Cancer	48

Asthma	47
Esophageal Cancer	40

B.2.2 More results of different stopping rules

Figures B.1 to B.5 show more detailed results for the proportion of correctly classified deaths among deaths due to each COD using different stopping criteria.

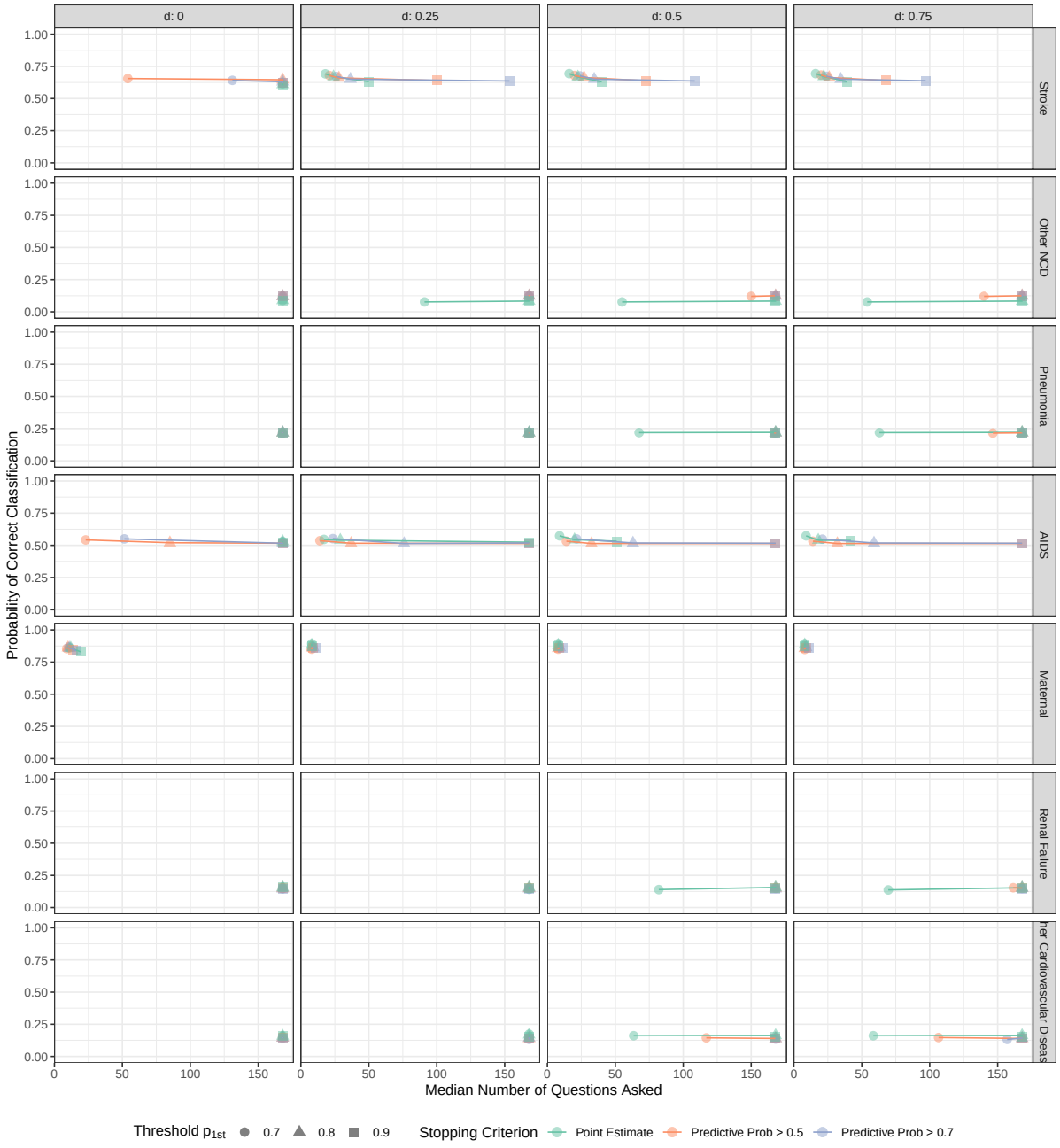


Figure B.1: Proportion of correctly classified deaths among deaths due to each cause using different stopping criteria. The causes are listed in order of their sample size in the PHMRC data.

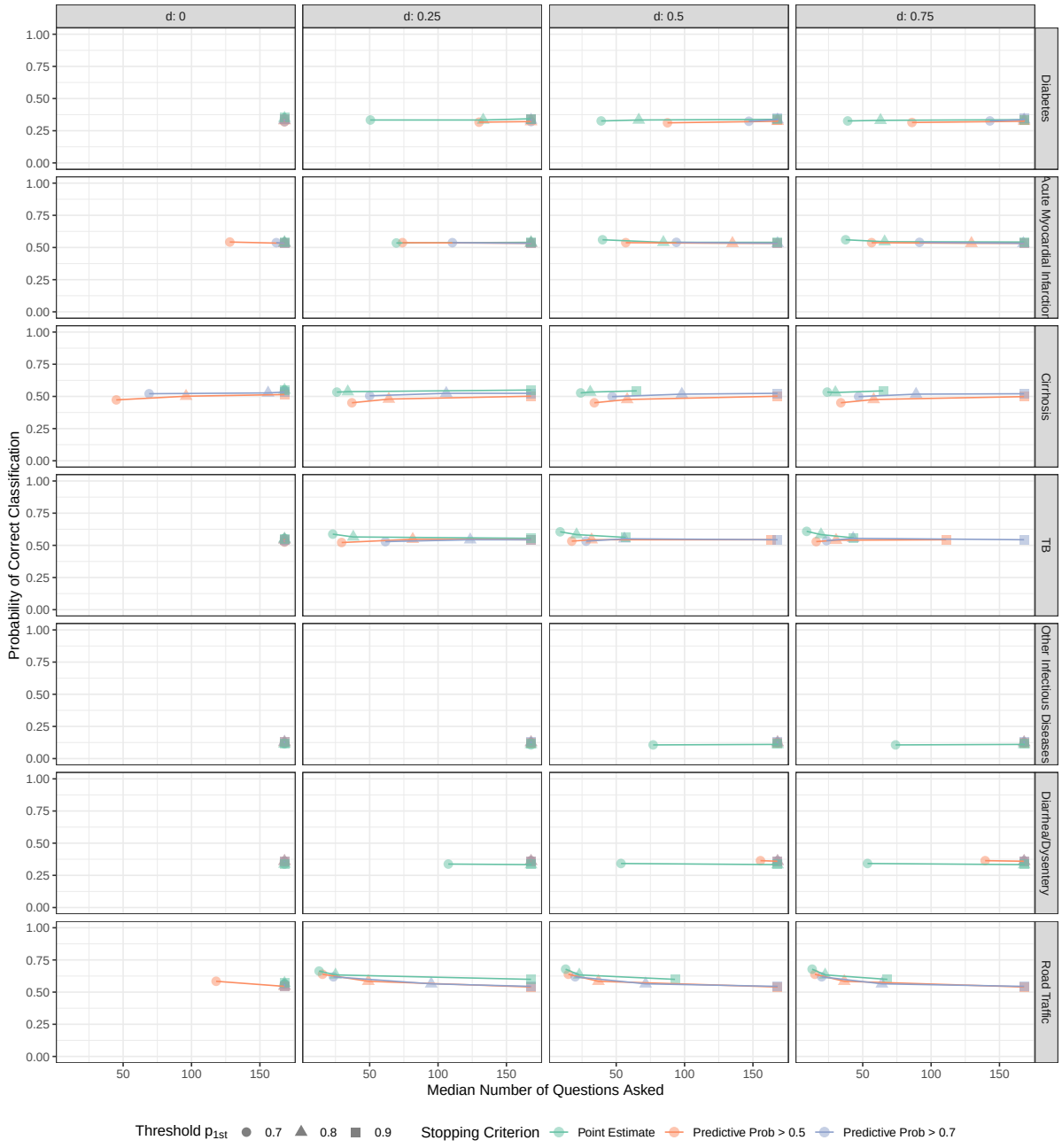


Figure B.2: (Continued) Proportion of correctly classified deaths among deaths due to each cause using different stopping criteria. The causes are listed in order of their sample size in the PHMRC data.

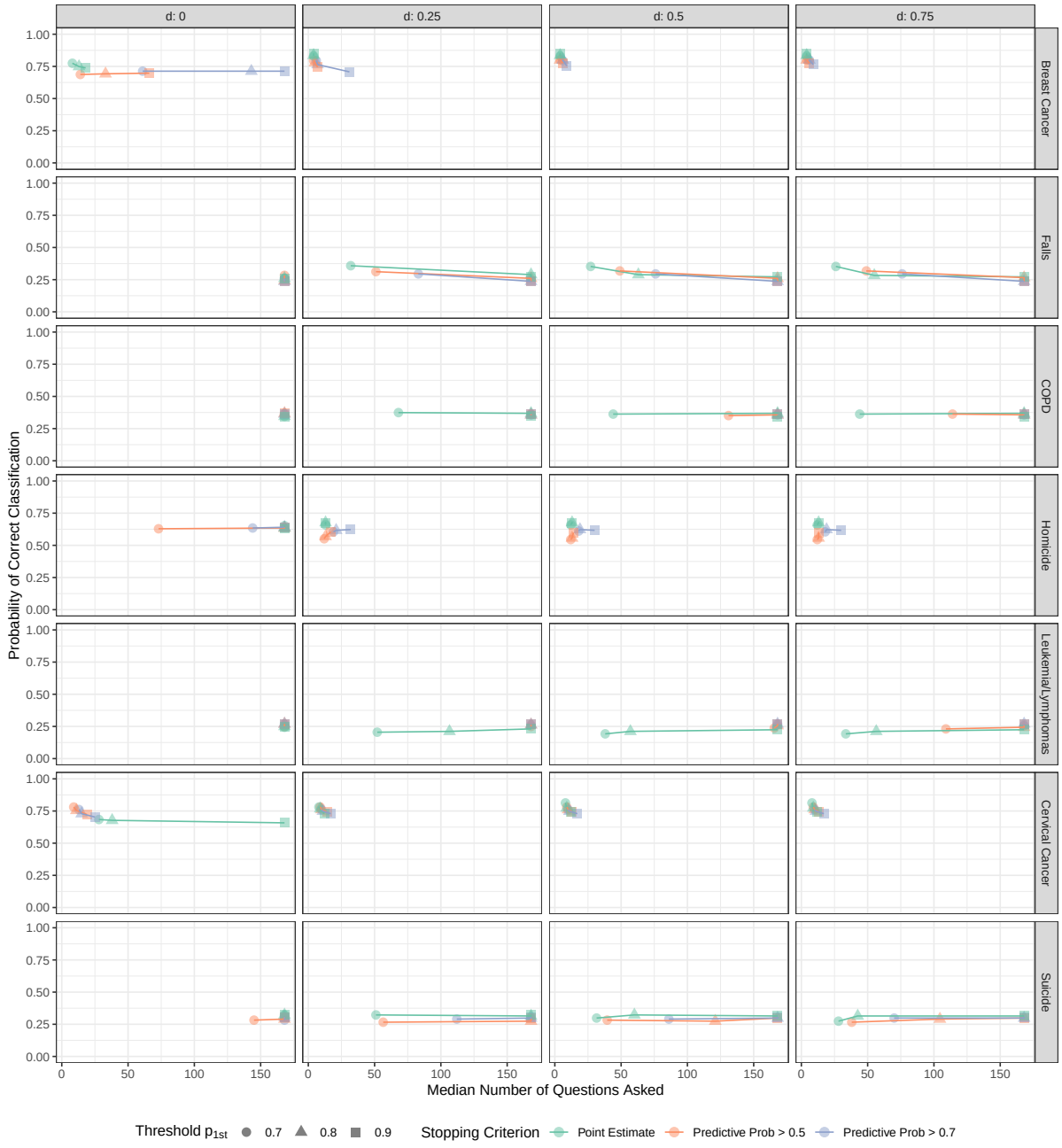


Figure B.3: (Continued) Proportion of correctly classified deaths among deaths due to each cause using different stopping criteria. The causes are listed in order of their sample size in the PHMRC data.

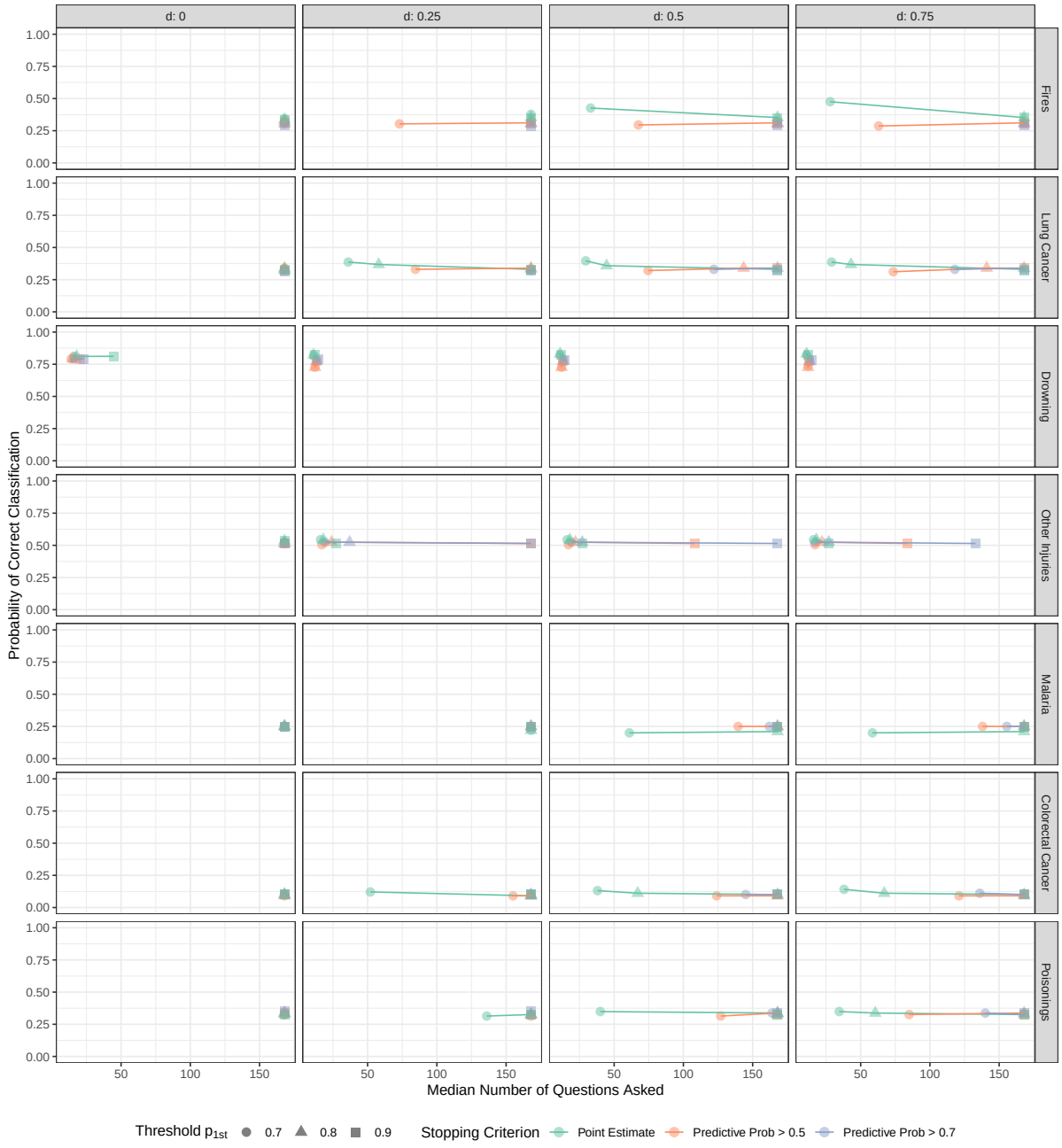


Figure B.4: (Continued) Proportion of correctly classified deaths among deaths due to each cause using different stopping criteria. The causes are listed in order of their sample size in the PHMRC data.

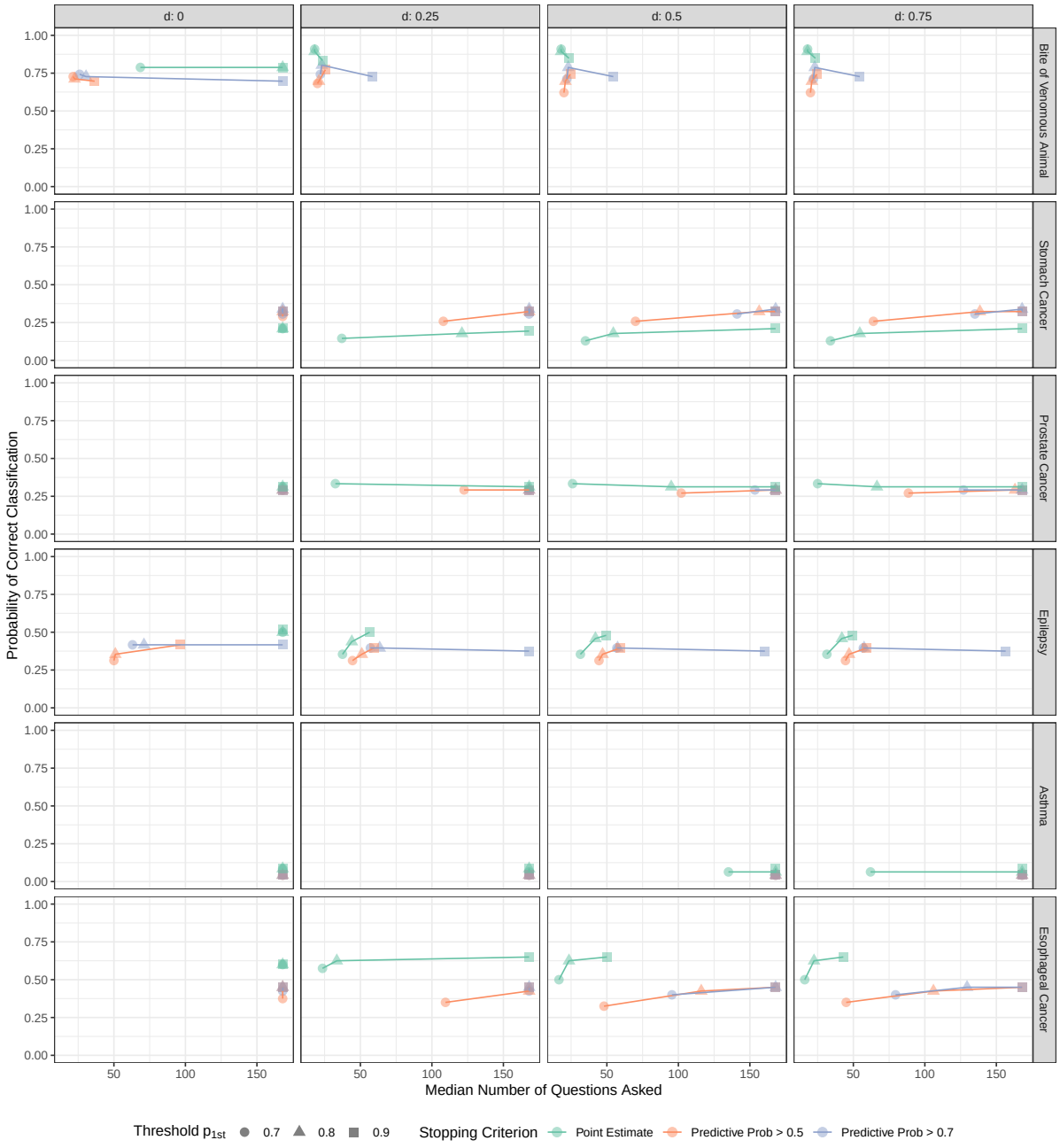


Figure B.5: (Continued) Proportion of correctly classified deaths among deaths due to each cause using different stopping criteria. The causes are listed in order of their sample size in the PHMRC data.

Appendix C

SUPPLEMENT TO CHAPTER 4

C.0.1 ICD-10 COD Classification

Mapping 34 PHMRC all-cause mortality labels to five Broad COD labels

ICD-10 Code	**All-Cause Mortality Label**	**Broad COD Label**
K74	cirrhosis	non-communicable
G40	epilepsy	non-communicable
J18	pneumonia	communicable
J44	copd	non-communicable
I21	acute myocardial infarction	non-communicable
X00-X09	fires	external
N17-N19	renal failure	non-communicable
C34	lung cancer	non-communicable
O00-O99	maternal	maternal
W65-W74	drowning	external
I00-I99	other cardiovascular diseases	non-communicable
B20-B24	aids	aids-tb
Varies	other non-communicable diseases	non-communicable
W00-W19	falls	external
V01-V99	road traffic	external
E10-E14	diabetes	non-communicable
A00-B99	other infectious diseases	communicable
A15-A19	tb	aids-tb
X60-X84	suicide	external
Varies	other injuries	external
C53	cervical cancer	non-communicable
I60-I69	stroke	non-communicable
B50-B54	malaria	communicable
J45	asthma	non-communicable
C18-C20	colorectal cancer	non-communicable
X85-Y09	homicide	external

A09	diarrhea/dysentery	communicable
C50	breast cancer	non-communicable
C81-C96	leukemia/lymphomas	non-communicable
X40-X49	poisonings	external
C61	prostate cancer	non-communicable
C15	esophageal cancer	non-communicable
C16	stomach cancer	non-communicable
T63	bite of venomous animal	external

C.0.2 `PPI` and `PPI++`: overview

We formalize our problem setting while reviewing the PPI `PPI` and `PPI++` approach (Angelopoulos et al., 2023a,b). The goal of `PPI` is to estimate a d-dimensional parameter $\boldsymbol{\theta}$ with n labeled data points $(X_{li}, Y_{li}) \overset{\text{iid}}{\sim} \mathbb{P}, i \in \{1, \ldots, n\}$, and N unlabeled data points $X_{ui} \overset{\text{iid}}{\sim} \mathbb{P}_X, i \in \{1, \ldots, N\}$.

Additionally, there is a black-box model f that uses the features to predict the outcomes based on labeled and unlabeled data, denoted as $\hat{Y}_l^f = (\hat{Y}_{l1}^f, \ldots, \hat{Y}_{ln}^f)$ and $\hat{Y}_u^f = (\hat{Y}_{u1}^f, \ldots, \hat{X}_{uN}^f)$. The prediction rule f is assumed to be independent of the observed data. While `PPI` performs inference on a correction term to incorporate these black-box predictions; `PPI++` improves the methodology to be more computationally efficient and develops a modification that adapts to the accuracy of the prediction rule f by introducing an additional parameter λ that interpolates between classical andPPI.

The starting point of `PPI` and `PPI++` approach is to define the population losses:

$$L(\theta) = \mathbb{E}[l_\theta(X, Y)], \text{ and}$$

$$L^f(\theta) = \mathbb{E}[l_\theta(X, \hat{Y}^f)].$$

It is recognized that

$$L^{f_l}(\theta) = \mathbb{E}[l_\theta(X_l, \hat{Y}_l^f)] = \mathbb{E}[l_\theta(X_u, \hat{Y}_u^f)] = L^{f_u}(\theta).$$

Therefore, the "rectified" loss is defined as:

$$L^{PPI}(\theta) = L_n(\theta) + L_N^{f_u}(\theta) - L_n^{f_l}(\theta)$$

where

$$L_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} l_\theta(X_{li}, Y_{li}),$$

$$L_n^{f_l}(\theta) = \frac{1}{n}\sum_{i=1}^{n} l_\theta(X_{li}, \hat{Y}_{li}^f)), \text{ and}$$

$$L_N^{f_u}(\theta) = \frac{1}{N}\sum_{i=1}^{N} l_\theta(X_{ui}, \hat{Y}_{ui}^f).$$

The rectified loss leads to the prediction-powered point estimate:

$$\hat{\theta}^{PPI} = \arg\min_{\theta \in \mathbb{R}^d} L^{PPI}(\theta).$$

`PPI` constructs the confidence set $C_\alpha^{PPI} = \{\theta \in \mathbb{R}^d \text{ s.t. } \mathbf{0} \in C_\alpha^\nabla(\theta)\}$ for true parameter θ^* by forming $(1-\alpha)$-confidence sets, $C_\alpha^\nabla(\theta)$, for the gradient of the loss $\nabla L^{PPI}(\theta)$. This approach works well for a fixed θ but suffers from computational issue when forming the confidence set for every $\theta \in \mathbb{R}^d$. `PPI++` derives the asymptotic normality about $\hat{\theta}$. This result allows the construction of $(1-\alpha)$-confidence intervals $C_\alpha^{PPI} = \{\hat{\theta}_j^{PPI} \pm z_{1-\alpha/2} \times \hat{\sigma}_j/\sqrt{n}\}$, where $\hat{\sigma}_j^2$ is a consistent estimate of the asymptotic variance of $\hat{\theta}_j^{PPI}$ and z_q is the q-quantile of standard normal distribution. Similar confidence intervals centered around $\hat{\theta}^{PPI}$ can be constructed when the inference is performed on the whole θ^* vector rather than a single coordinate θ_j^*, which results in a far more efficient algorithm.

Moreover, `PPI++` generalizes `PPI` to allow the inference to adapt to the accuracy of the supplied predictions, yielding estimations that are never worse than the classical inference. This approach is called "power tuning" by incorporating a tuning parameter $\lambda \in [0, 1]$ in the rectified loss, leading to the following estimator:

$$\hat{\theta}_\lambda^{\texttt{PPI++}} = \arg\min_\theta L_\lambda^{\texttt{PPI++}}(\theta),$$

where

$$L_\lambda^{\texttt{PPI++}}(\theta) = L_n(\theta) + \lambda(L_N^{f_u}(\theta) - L_n^{f_l}(\theta)).$$

`PPI++` recovers the original `PPI` approach when $\lambda = 1$ and reduces to classical inference when $\lambda = 0$. One can adaptively choose a data-dependent tuning parameter $\hat{\lambda}$ to maximize the estimation and inferential efficiency.

Algorithm 1: `multiPPI++`

Input: labeled K-category COD data $\{(X_{li}, Y_{li})\}_{i=1}^n$, unlabeled data $\{X_{ui}\}_{i=1}^N$, NLP model f, significance level $\alpha \in (0,1)$, coefficient index $j \in [d(K-1)]$

1. Optimally select tuning parameter $\hat{\lambda}$ `// set tuning parameter`

2. $\hat{\theta}_{\hat{\lambda}}^{\mathtt{m}} = \underset{\theta \in \mathbb{R}^{d(K-1)}}{\arg\min}\ L_{\hat{\lambda}}^{\mathtt{m}}(\theta)$ `// multiPPI++ estimator`

3. $\hat{H} = \frac{1}{N+n}(\sum_{i=1}^n \psi''(X_{li}^T \hat{\theta}_{\hat{\lambda}}^{\mathtt{m}}) X_{li} X_{li}^T + \sum_{i=1}^N \psi''(X_{ui}^T \hat{\theta}_{\hat{\lambda}}^{\mathtt{m}}) X_{ui} X_{ui}^T)$, where $\psi(\theta, x) = \log\left(\sum_{k=1}^{K-1} e^{x^T \theta_k}\right), \theta_k \in \mathbb{R}^d$ `// empirical Hessian`

4. $\hat{\Sigma} = \hat{H}^{-1}(\frac{n}{N}\hat{V}_f + \hat{V}_\Delta)\hat{H}^{-1}$, where

   $\hat{V}_f = \hat{\lambda}^2 \widehat{\mathrm{Cov}}_{N+n}((\psi'(X_{li}^T \hat{\theta}_{\hat{\lambda}}^{\mathtt{m}}) - \hat{Y}_{li}^f))X_{li})$ and
   $\hat{V}_\Delta = \widehat{\mathrm{Cov}}_n((1-\hat{\lambda})(\psi'(X_{li}^T \hat{\theta}_{\hat{\lambda}}^{\mathtt{m}}) + (\hat{\lambda}\hat{Y}_{li}^f - Y_{li})X_{li})$
   `// covariance estimator`

Output:

Prediction-powered point estimates $\hat{\theta}_{\hat{\lambda}}^{\mathtt{m}}$ and confidence interval
$\mathcal{C}_\alpha^{\mathtt{m}} = \left(\hat{\theta}_{\hat{\lambda},j}^{\mathtt{m}} \pm z_{1-\alpha/2}\sqrt{\hat{\Sigma}_{jj}/n}\right)$ for coordinate j

---

C.0.3 `multiPPI++`

In K-class multinomial classification problem with outcomes $Y_i \in \{0, \ldots, K-1\}$ and covariates $X_i \in \mathbb{R}^d$, the parameter of interests θ is a $d(K-1)$-dimensional vector whose k-th block $\theta_k \in \mathbb{R}^d$ of components represent parameters for class $k \in \{1, \ldots, K-1\}$ excluding the reference class to avoid overparameterization. The loss function in classical inference takes the form

$$l_\theta(X, Y) = -\frac{1}{n}\sum_{i=1}^{n} X_i^T \theta_{Y_i} + \log\left(\sum_{k=1}^{K-1} e^{X_i^T \theta_k}\right),$$

Defining

$$\psi(\theta, x) = \log\left(\sum_{k=1}^{K-1} e^{x^T \theta_k}\right),$$

and the `multiPPI++` loss is given by

$$\begin{aligned} L_\lambda^{\mathtt{m}}(\theta) &= L_n(\theta) + \lambda(L_N^{f_u}(\theta) - L_n^{f_l}(\theta)) \\ &= -\frac{1}{n}\sum_{i=1}^{n}(X_{li}\theta_{Y_{li}} - \psi(\theta, X_{li})) \\ &\quad - \frac{\lambda}{N}\sum_{i=1}^{N}(X_{ui}^T \theta_{\hat{Y}_{ui}^f} - \psi(\theta, X_{ui})) + \frac{\lambda}{n}\sum_{i=1}^{n}(X_{li}^T \theta_{\hat{Y}_{li}^f} - \psi(\theta, X_{li})) \end{aligned}$$

Angelopoulos et al. (2023b) shows that the parameter λ can be optimally tuned to minimize the asymptotic variance of the prediction-powered estimate, leading to better estimation and inference than both the classical and `PPI` strategies. Particularly, in finite samples, we can obtain the plug-in estimate.

$$\hat{\lambda} = \frac{1}{2(1+\frac{n}{N})} \times \frac{\mathrm{Tr}\left(\widehat{H}_{\hat{\theta}_{PPI}}^{-1}\left(\widehat{\mathrm{Cov}}_n(\nabla l_{\hat{\theta}_{PPI}}, \nabla l_{\hat{\theta}_{PPI}}^f) + \widehat{\mathrm{Cov}}_n(\nabla l_{\hat{\theta}_{PPI}}^f, \nabla l_{\hat{\theta}_{PPI}})\right)\widehat{H}_{\hat{\theta}_{PPI}}^{-1}\right)}{\mathrm{Tr}\left(\widehat{H}_{\hat{\theta}_{PPI}}^{-1}\widehat{\mathrm{Cov}}_{N+n}(\nabla l_{\hat{\theta}_{PPI}}^f)\widehat{H}_{\hat{\theta}_{PPI}}^{-1}\right)}$$

where $\hat{\theta}_{PPI} = \hat{\theta}_\lambda^{\mathtt{PPI++}}$ is obtained by taking a fixed $\lambda \in [0, 1]$.

We summarize the `multiPPI++` adjusted inference procedures for multinomial logistic regression coefficients on predicted CODs by an NLP model f against covariate X, under significance level $\alpha \in (0, 1)$ in Algorithm 1.

C.0.4 Parameter Estimates Across All Sites: Full Data 80/20 Split

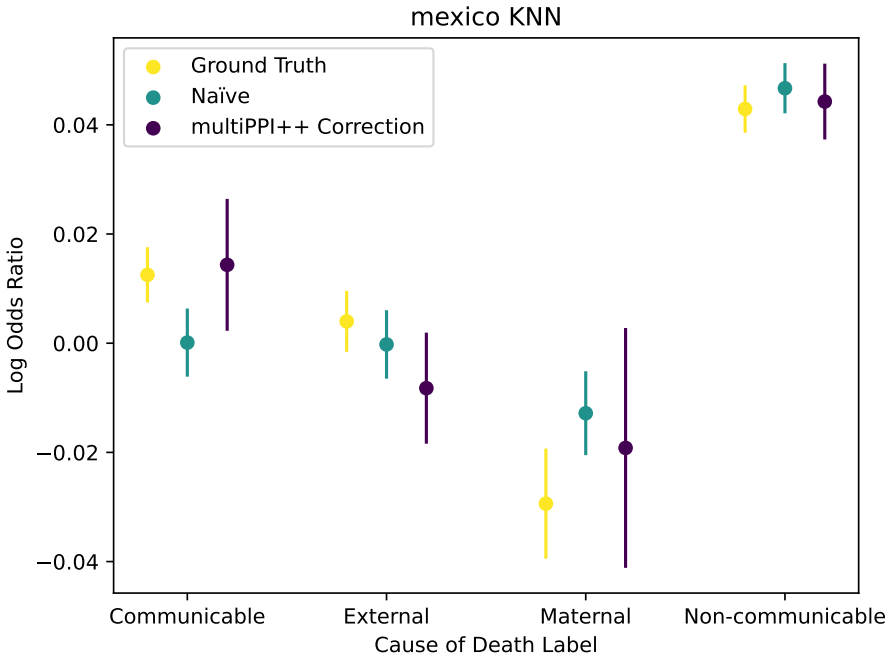


Figure C.1: Full Data 80/20 Split: Site/Model 1

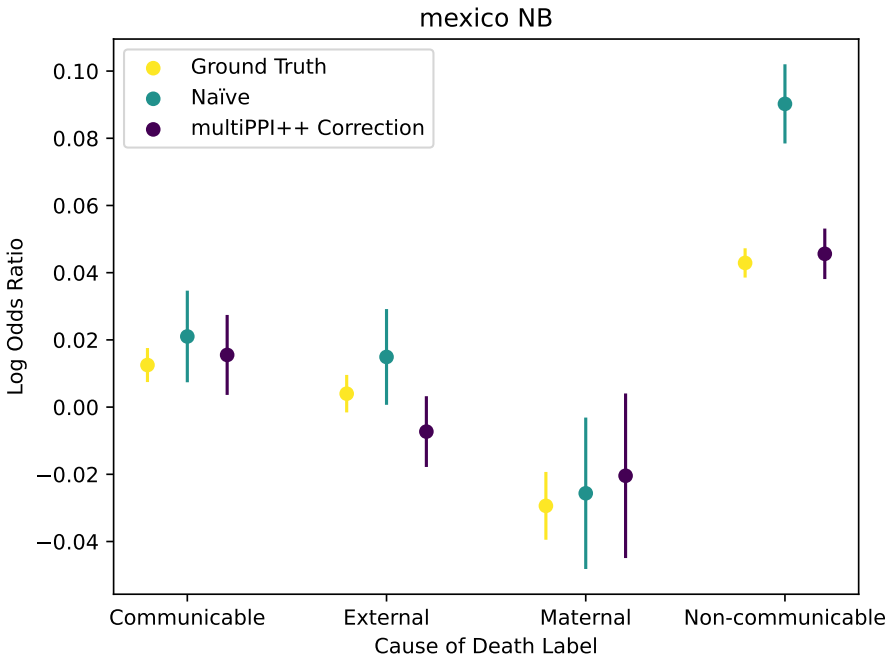


Figure C.2: Full Data 80/20 Split: Site/Model 2

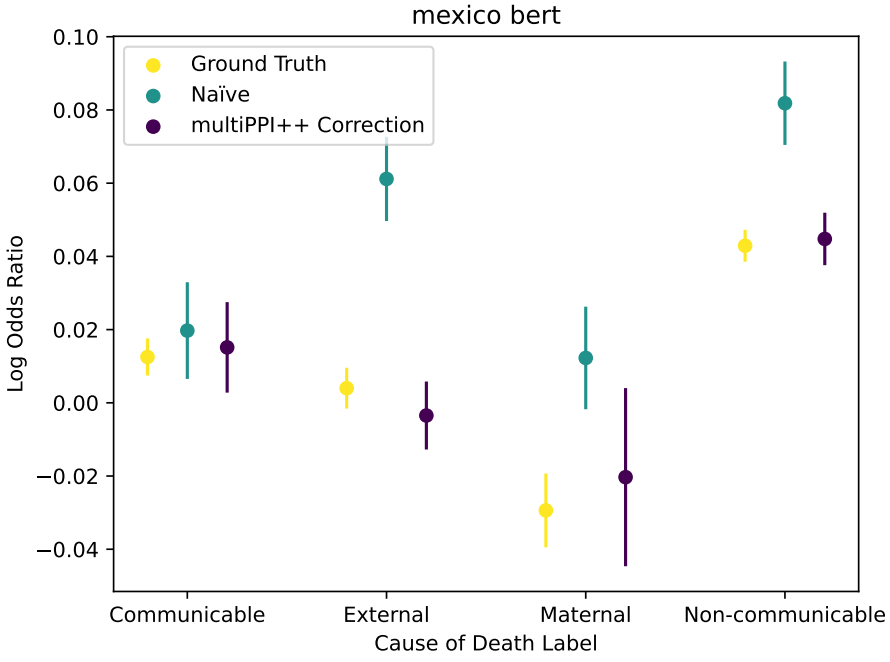


Figure C.3: Full Data 80/20 Split: Site/Model 3

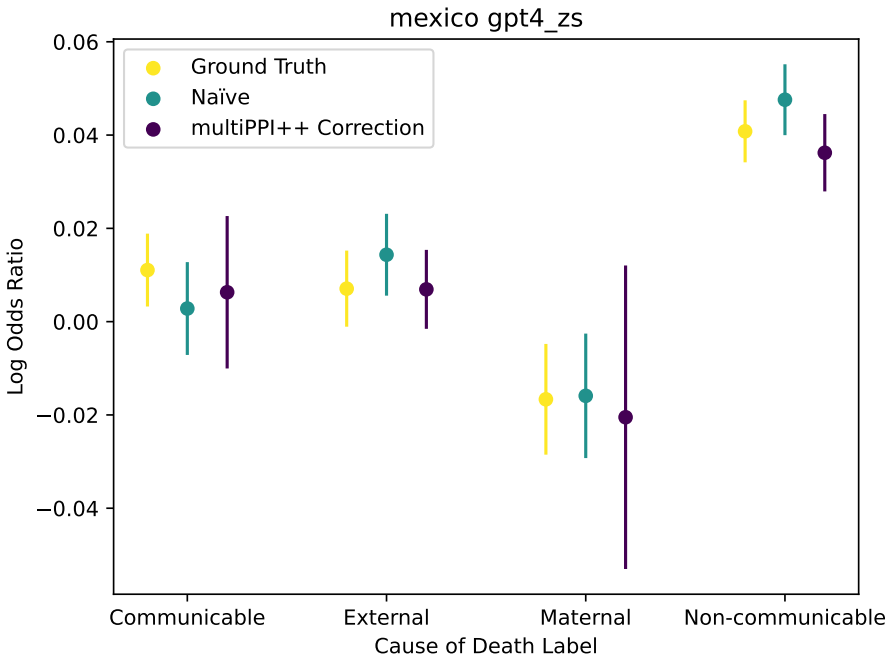


Figure C.4: Full Data 80/20 Split: Site/Model 4

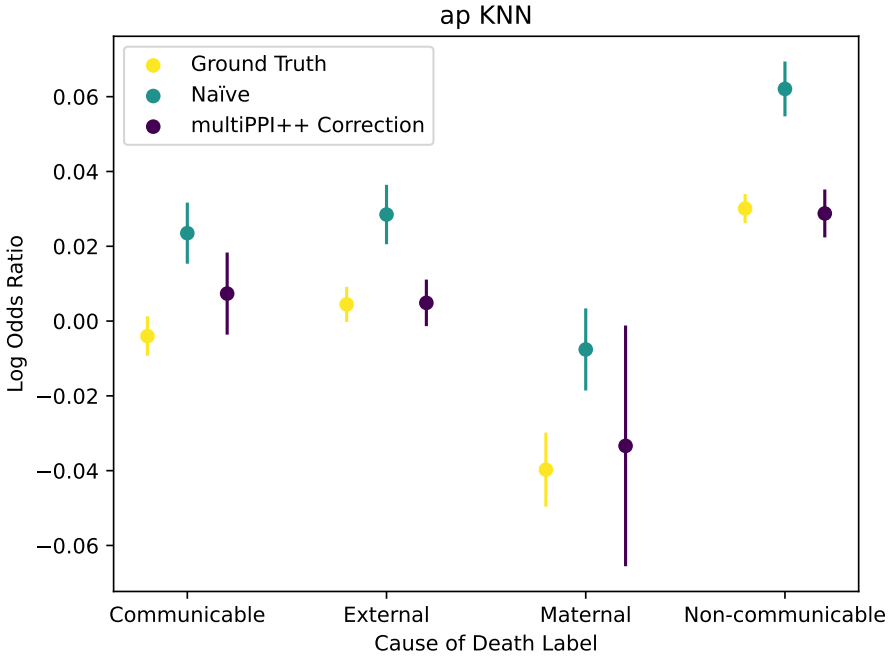


Figure C.5: Full Data 80/20 Split: Site/Model 5

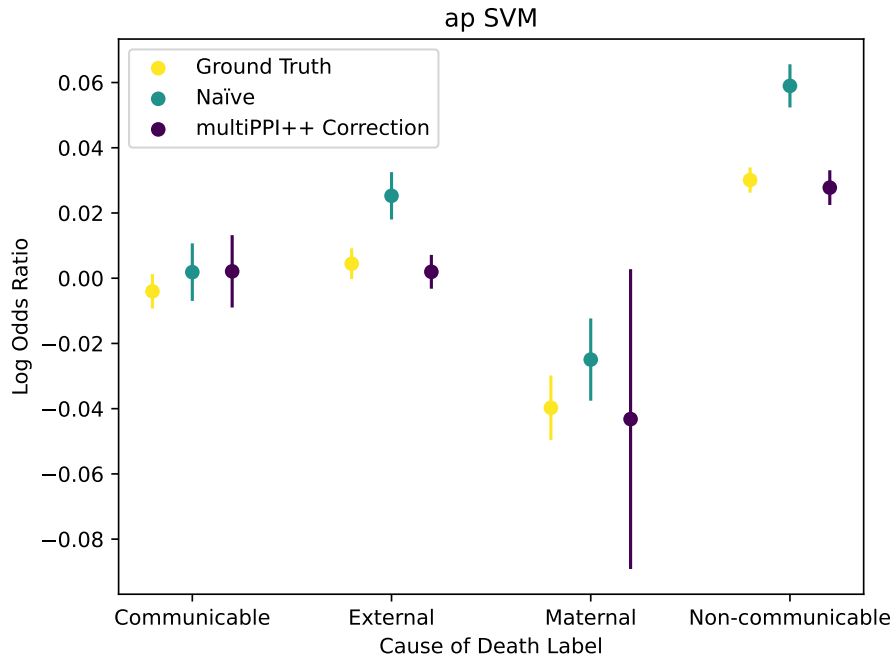


Figure C.6: Full Data 80/20 Split: Site/Model 6

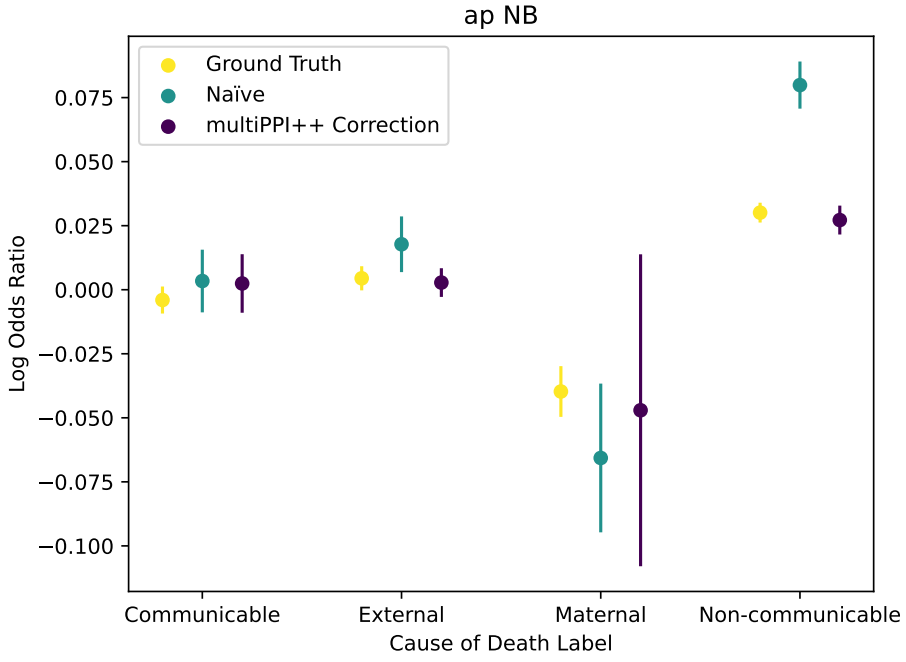


Figure C.7: Full Data 80/20 Split: Site/Model 7

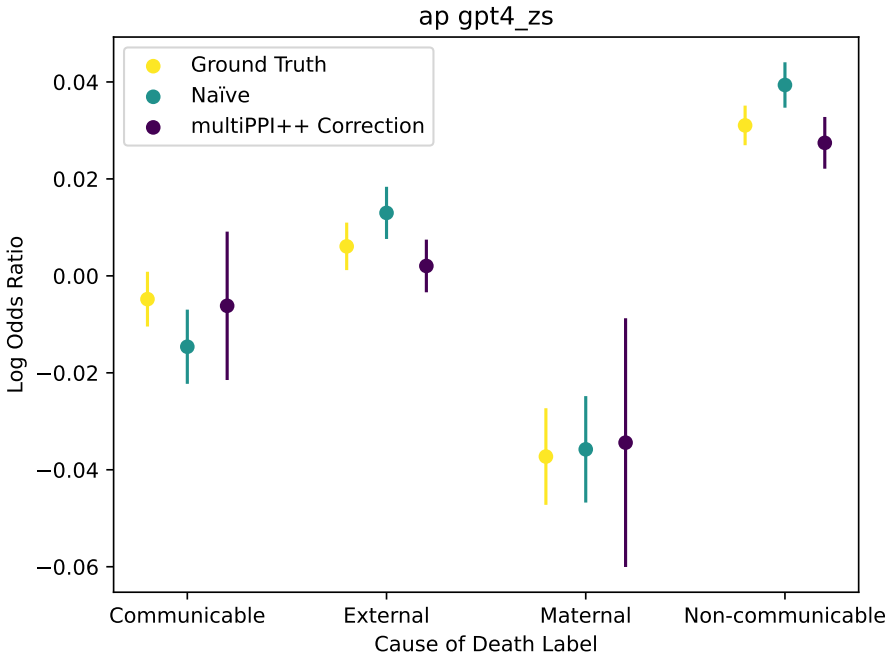


Figure C.8: Full Data 80/20 Split: Site/Model 8

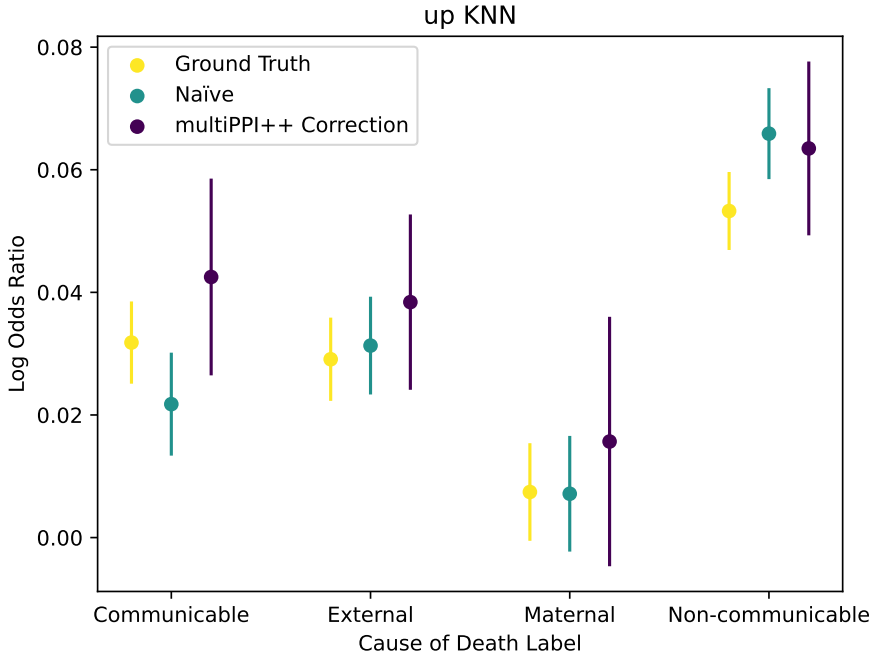


Figure C.9: Full Data 80/20 Split: Site/Model 9

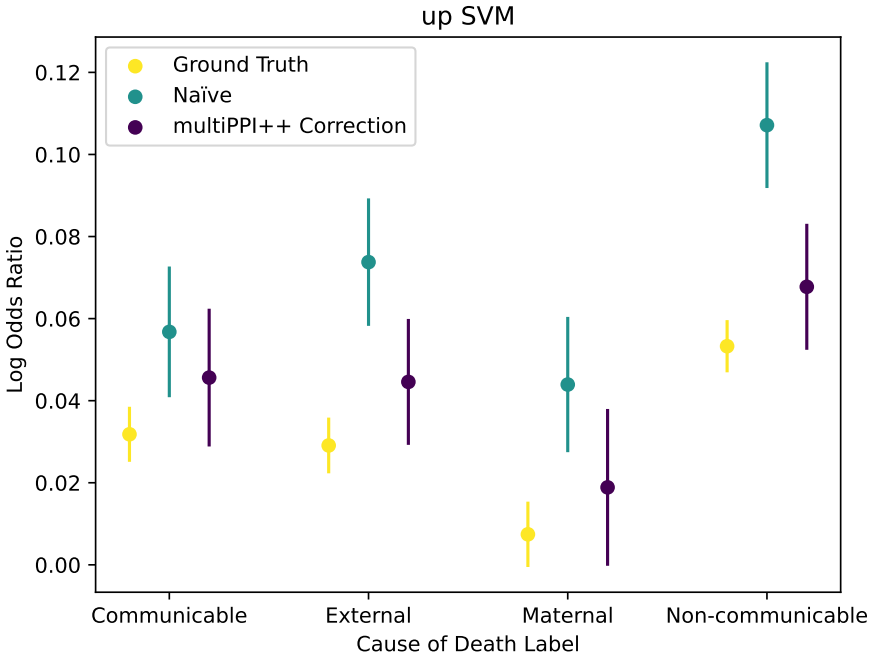


Figure C.10: Full Data 80/20 Split: Site/Model 10

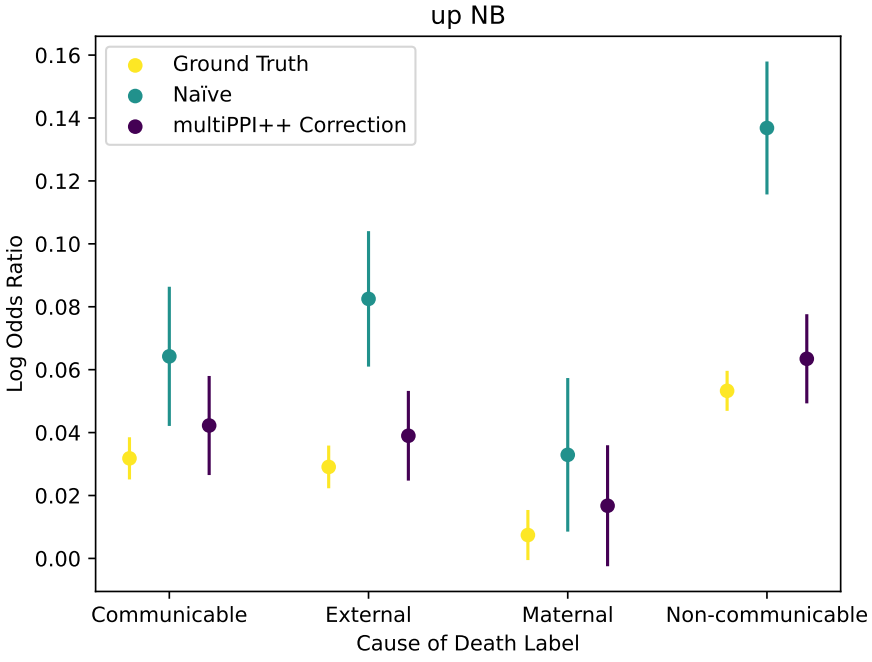


Figure C.11: Full Data 80/20 Split: Site/Model 11

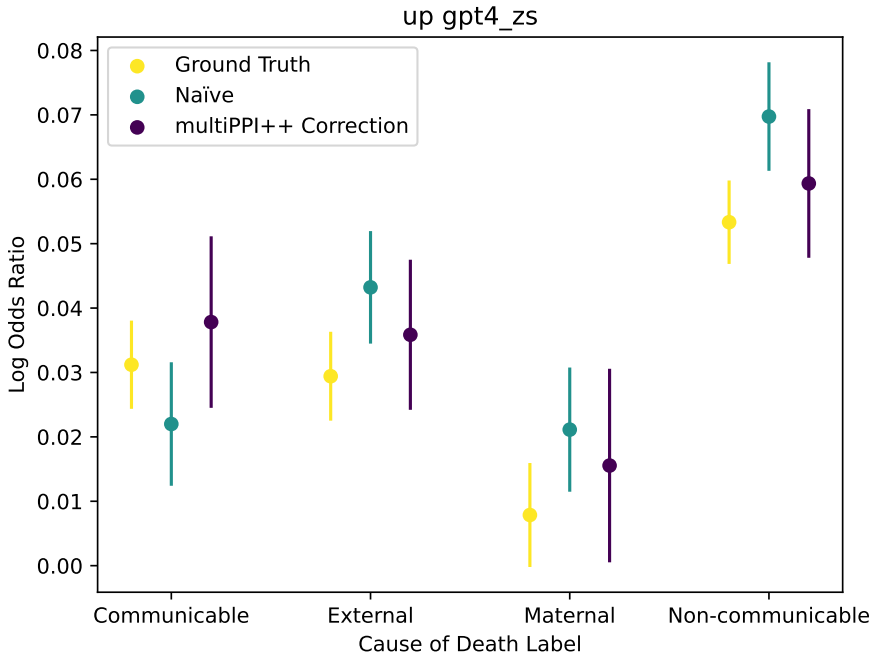


Figure C.12: Full Data 80/20 Split: Site/Model 12

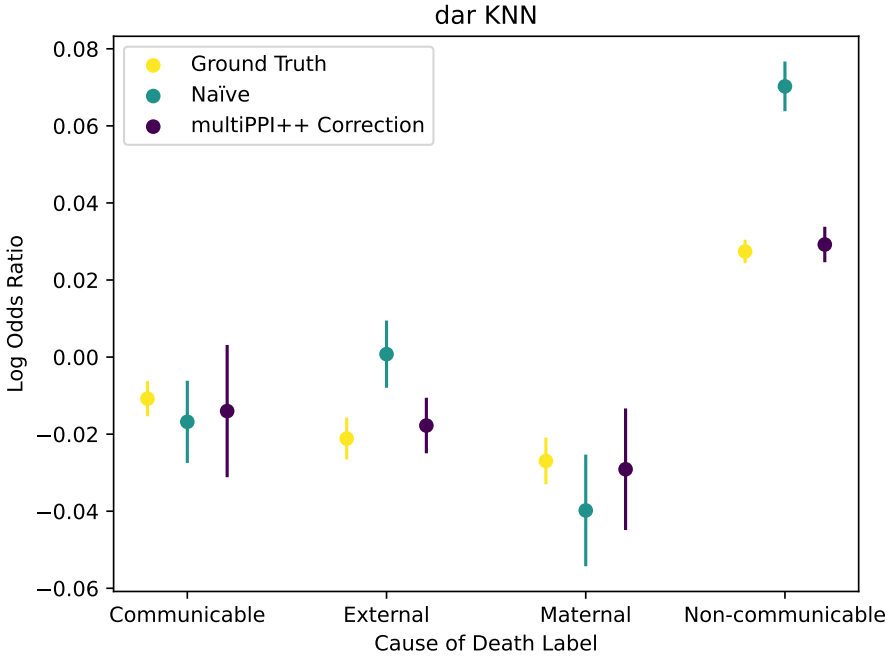


Figure C.13: Full Data 80/20 Split: Site/Model 13

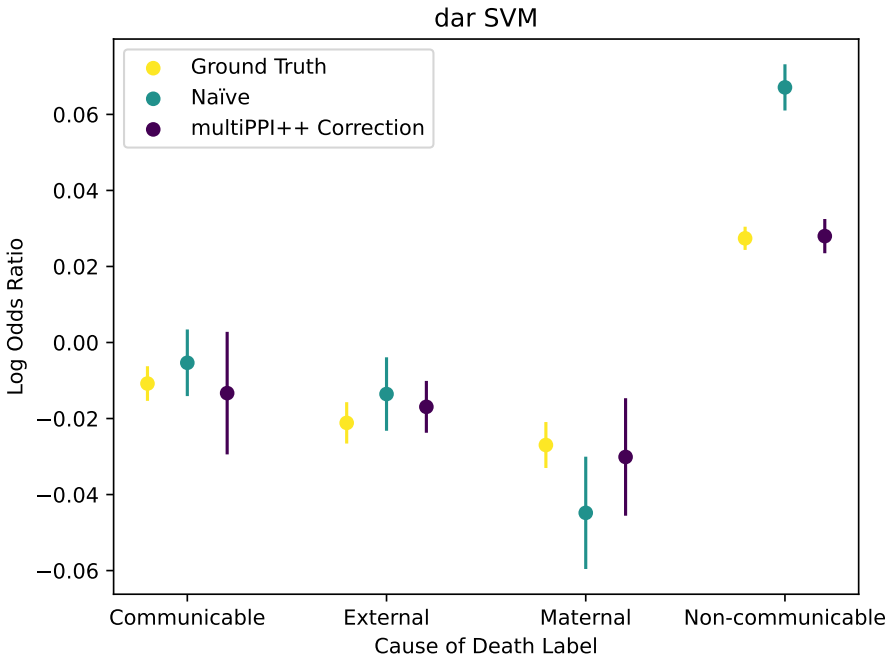


Figure C.14: Full Data 80/20 Split: Site/Model 14

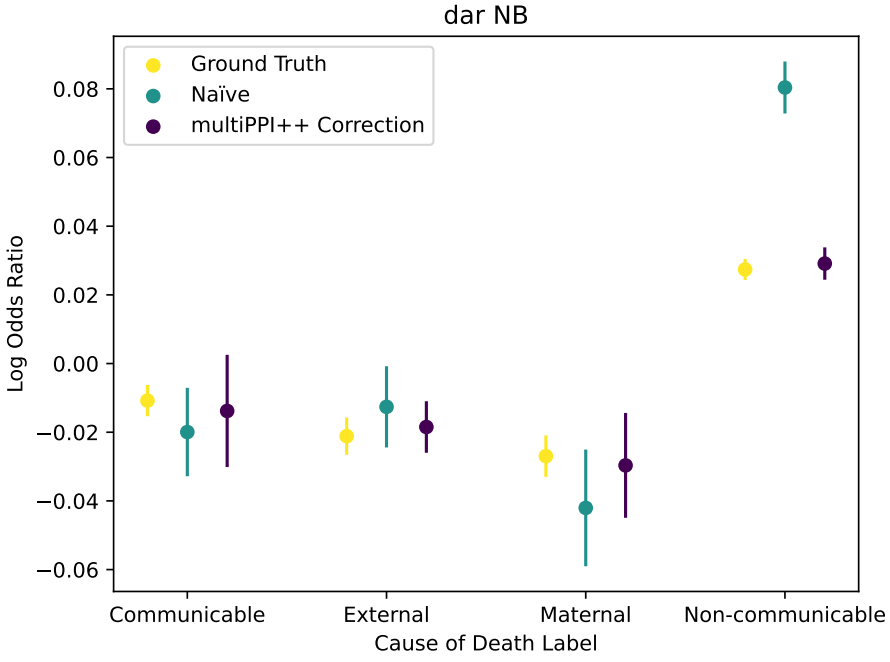


Figure C.15: Full Data 80/20 Split: Site/Model 15

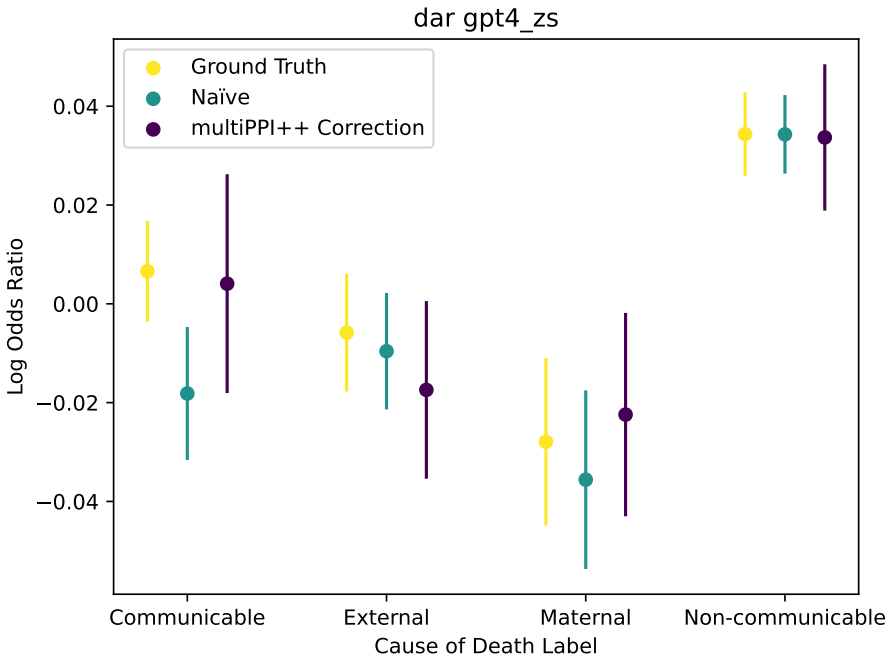


Figure C.16: Full Data 80/20 Split: Site/Model 16

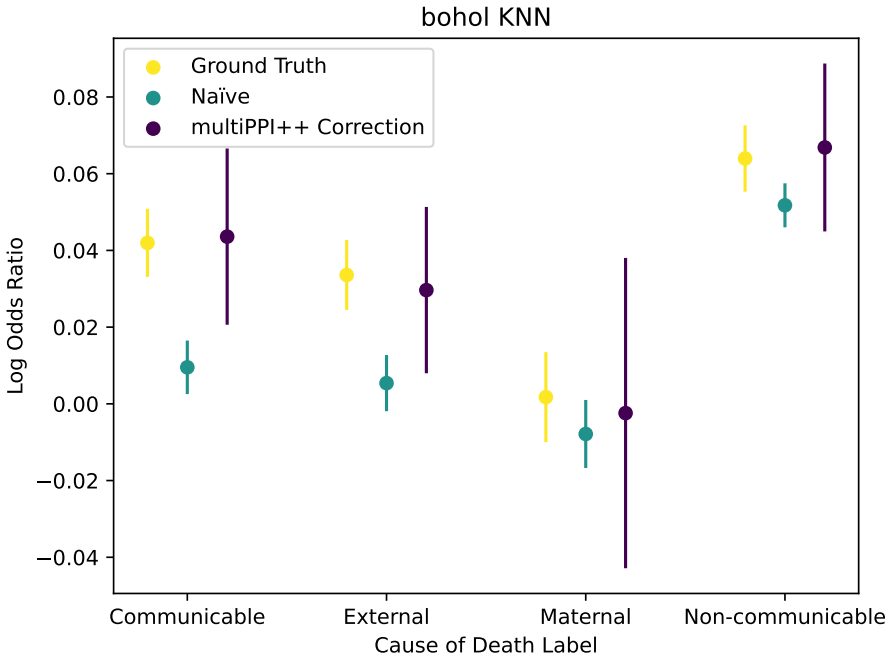


Figure C.17: Full Data 80/20 Split: Site/Model 17

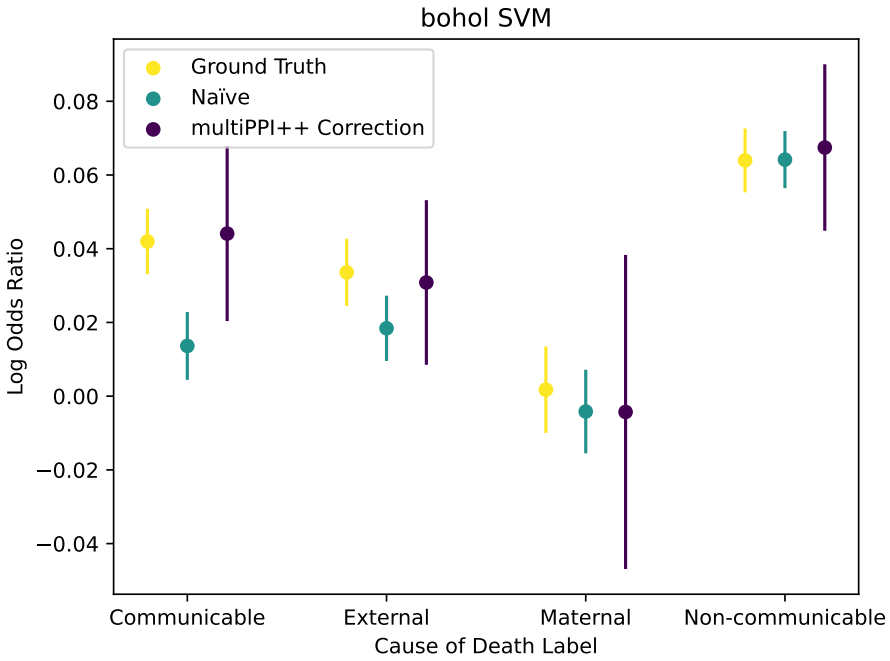


Figure C.18: Full Data 80/20 Split: Site/Model 18

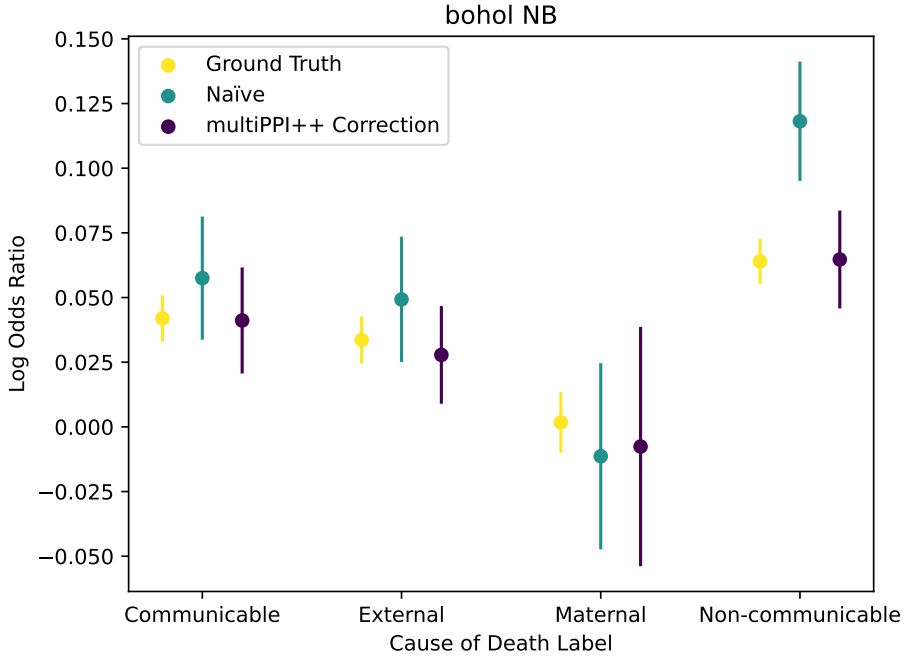


Figure C.19: Full Data 80/20 Split: Site/Model 19

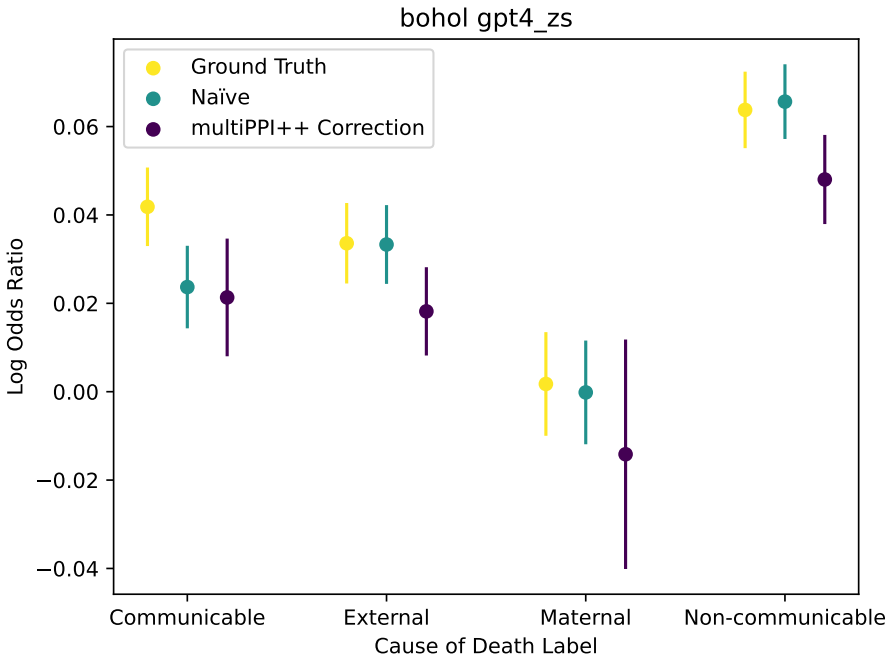


Figure C.20: Full Data 80/20 Split: Site/Model 20

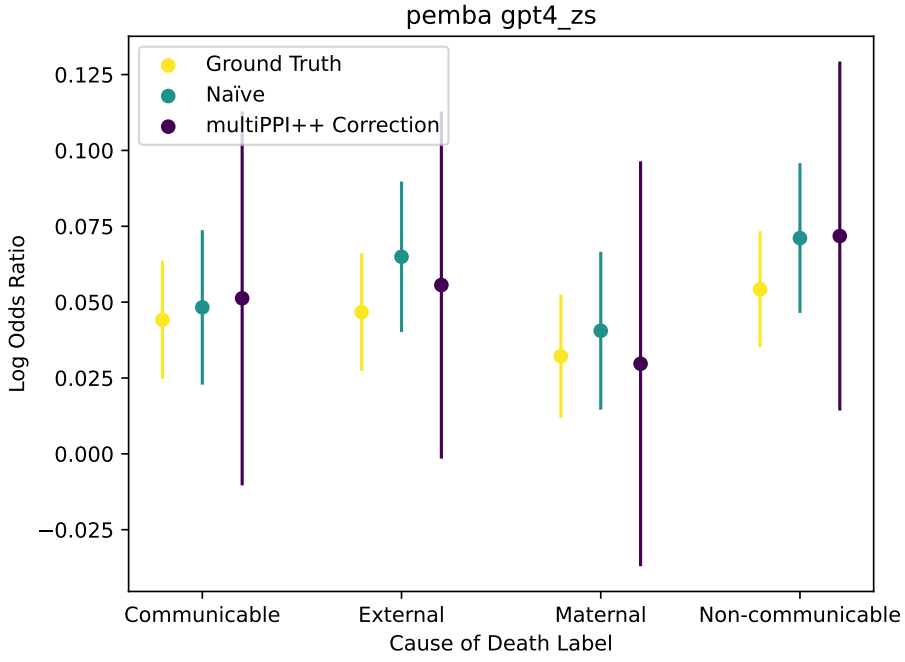


Figure C.21: Full Data 80/20 Split: Site/Model 21

C.0.5 Sensitivity Analysis

In a sensitivity analysis, we varied our data splitting strategy, which revealed nuanced insights into the classifier's behavior. Notably, when maintaining the split proportions (80/20) within each COD, the `PPI++` classifier exhibited minimal utilization of the labeled data, as indicated by the relatively small values of λ. Conversely, when employing an 80/20 split on the entire dataset, this led to a significant information loss for the minority classes, resulting in larger λ values for inevitable splits. Since an (80/20) on the whole dataset more closely resembles what one would see in reality in a prospective study. In contrast, an (80/20) split by COD resembles a retrospective study; this illustrates the importance of purposefully choosing an appropriate data splitting strategy to match the desired analysis.

C.0.6 Parameter Estimates Across All Sites: Stratified 80/20 Split

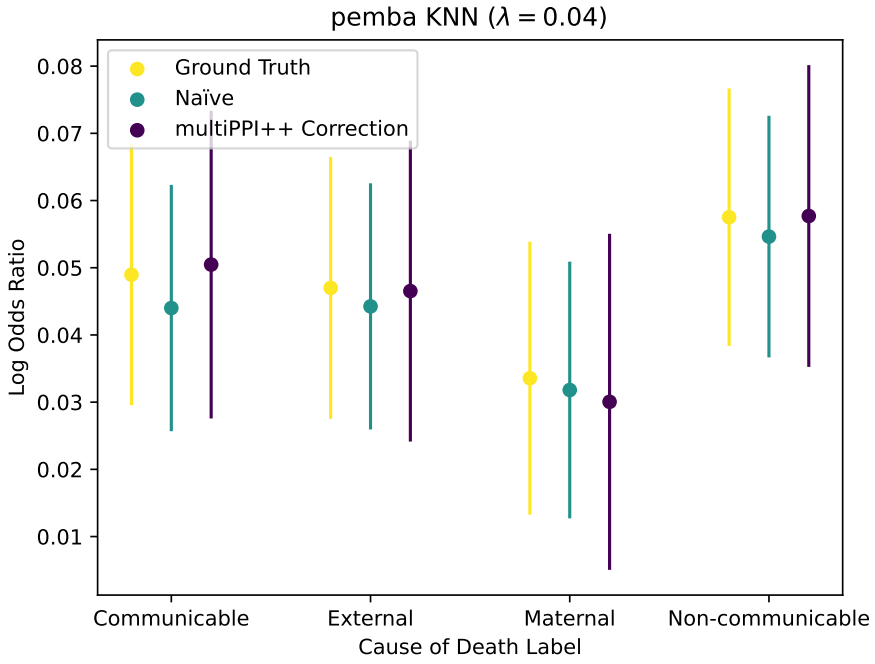


Figure C.22: Stratified 80/20 Split: Site/Model 22

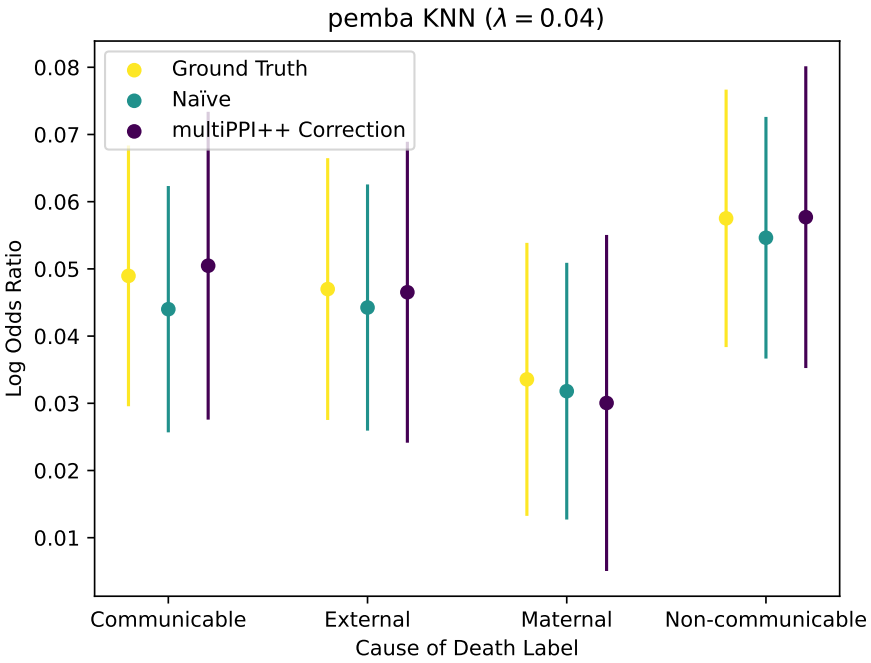


Figure C.23: Stratified 80/20 Split: Site/Model 22

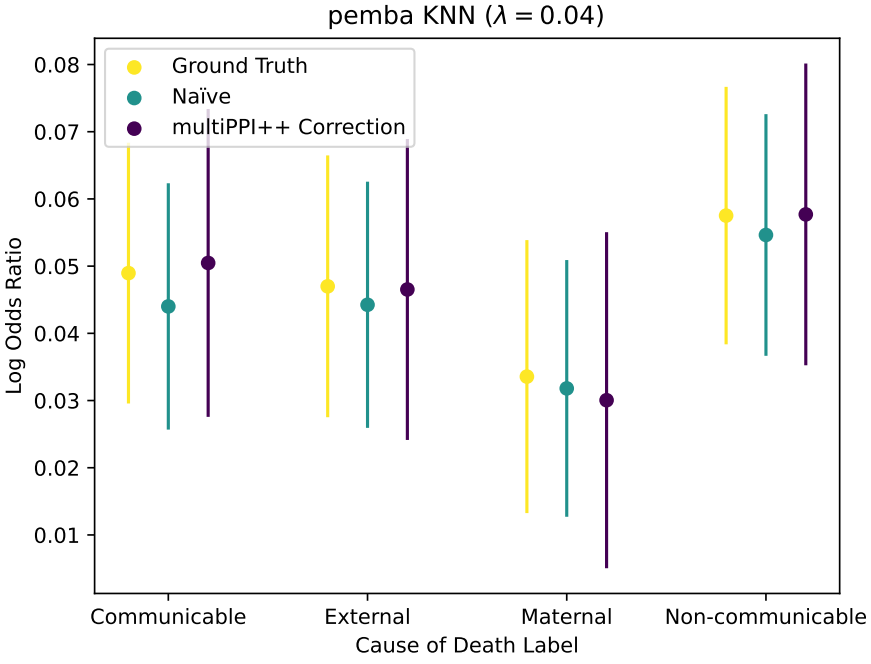


Figure C.24: Stratified 80/20 Split: Site/Model 22

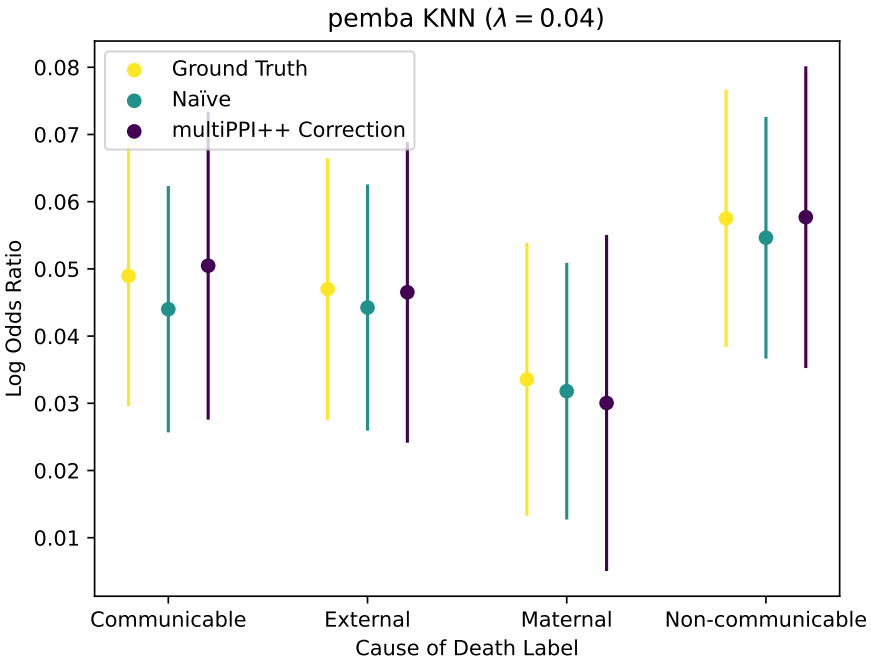


Figure C.25: Stratified 80/20 Split: Site/Model 22

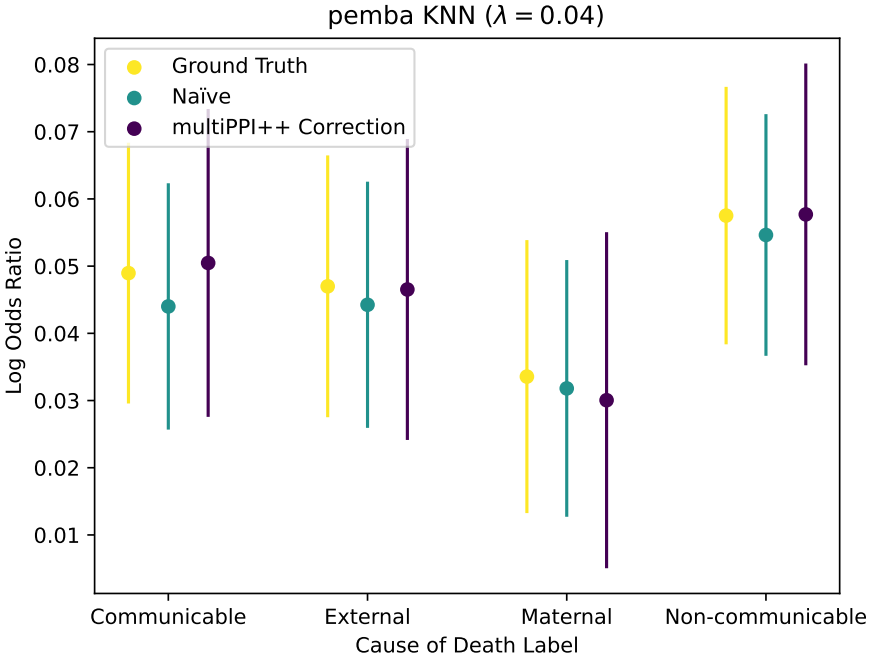


Figure C.26: Stratified 80/20 Split: Site/Model 22

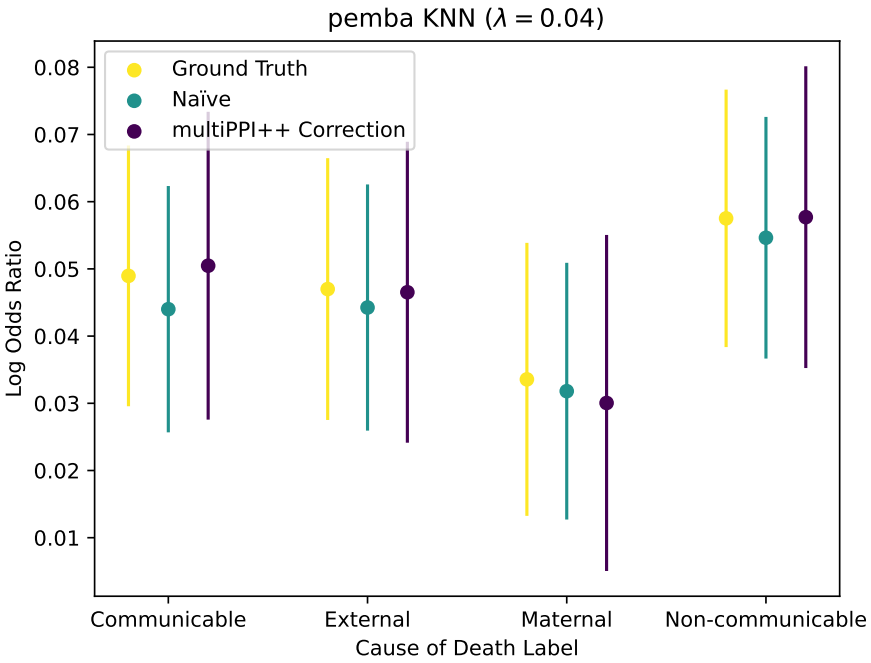


Figure C.27: Stratified 80/20 Split: Site/Model 22

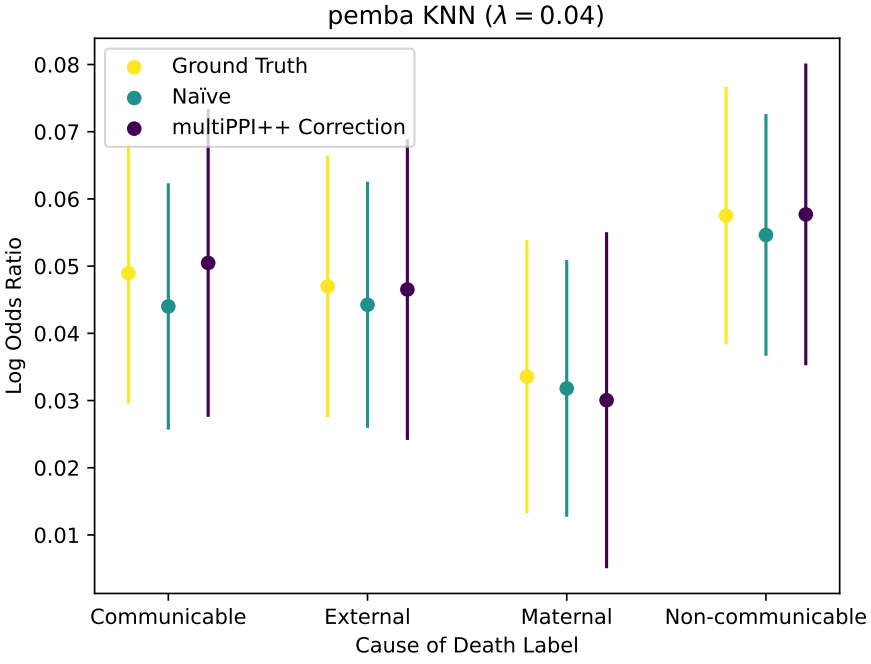


Figure C.28: Stratified 80/20 Split: Site/Model 22

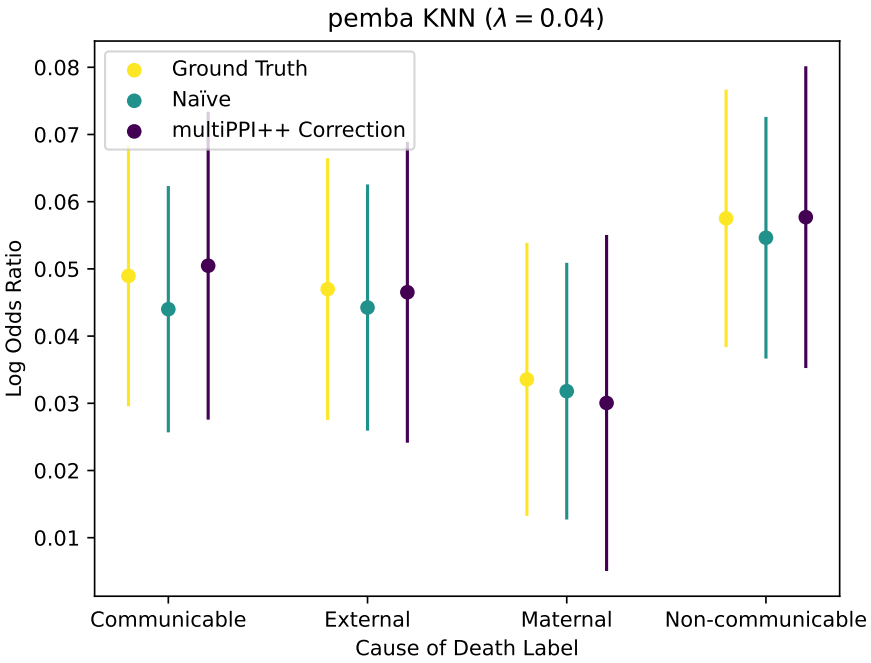


Figure C.29: Stratified 80/20 Split: Site/Model 22

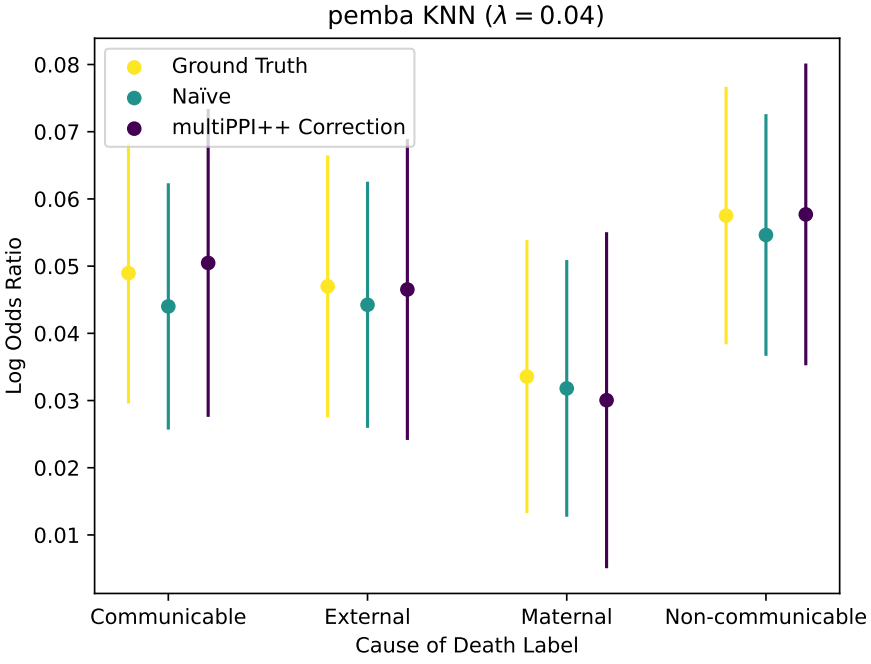


Figure C.30: Stratified 80/20 Split: Site/Model 22

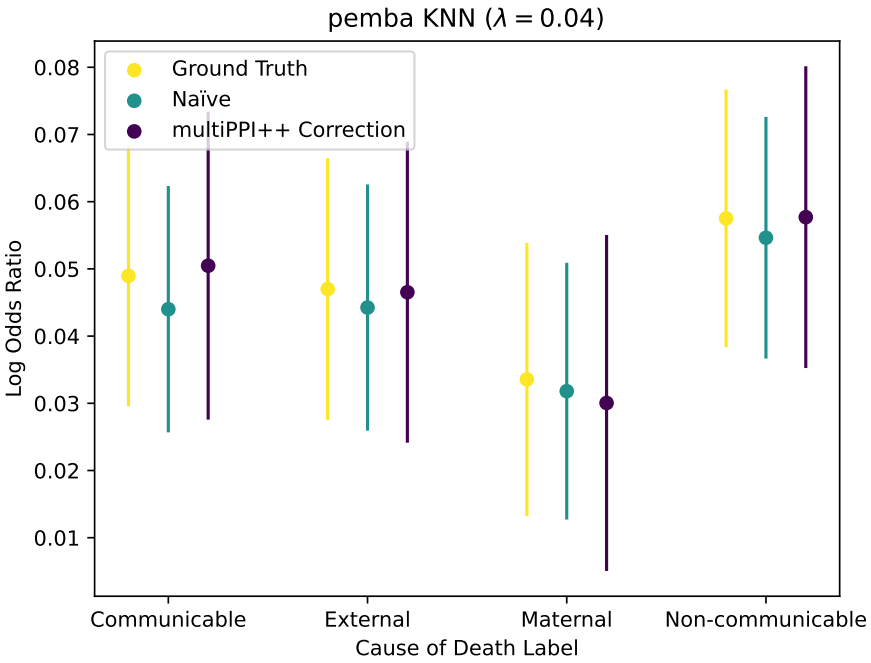


Figure C.31: Stratified 80/20 Split: Site/Model 22

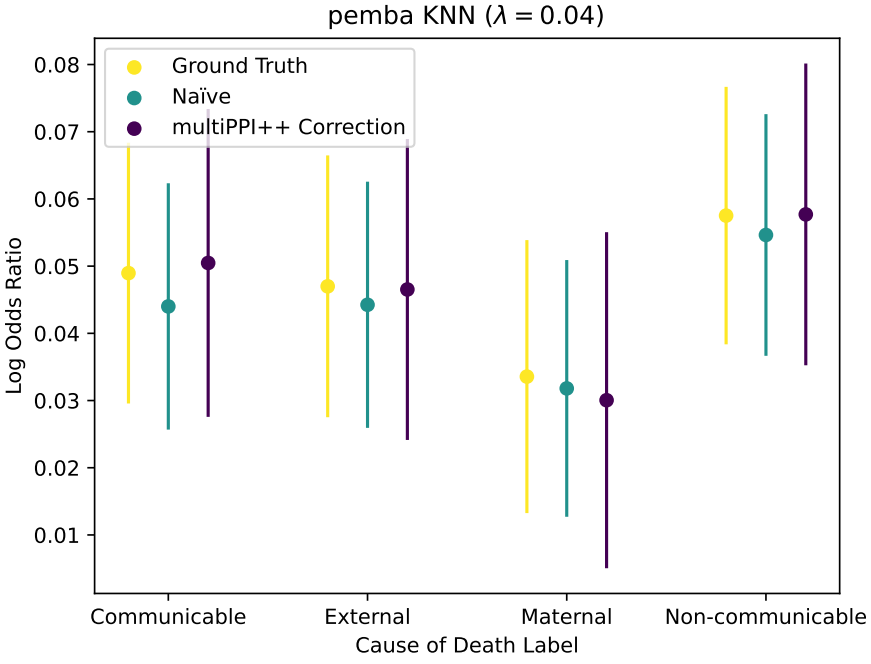


Figure C.32: Stratified 80/20 Split: Site/Model 22

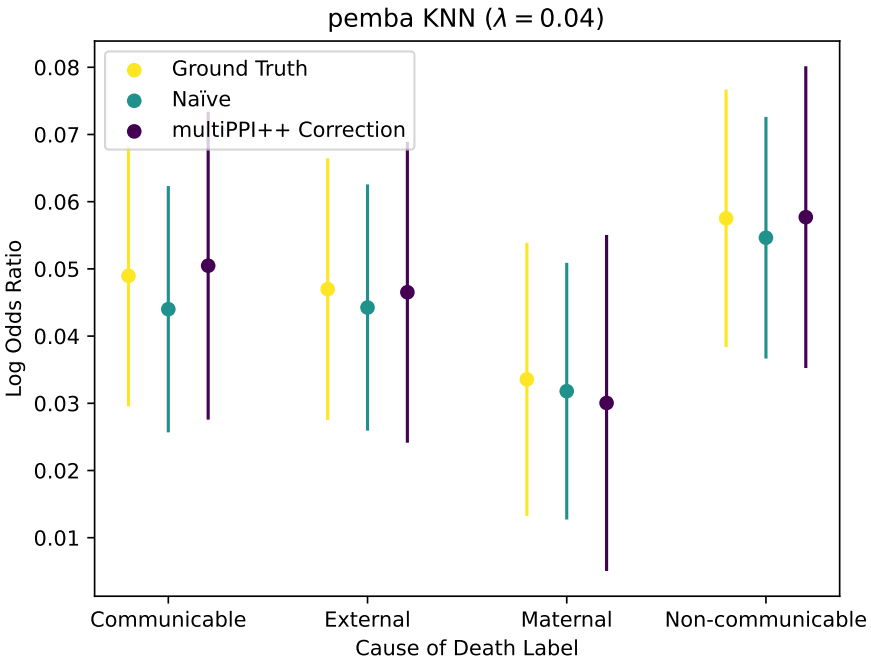


Figure C.33: Stratified 80/20 Split: Site/Model 22

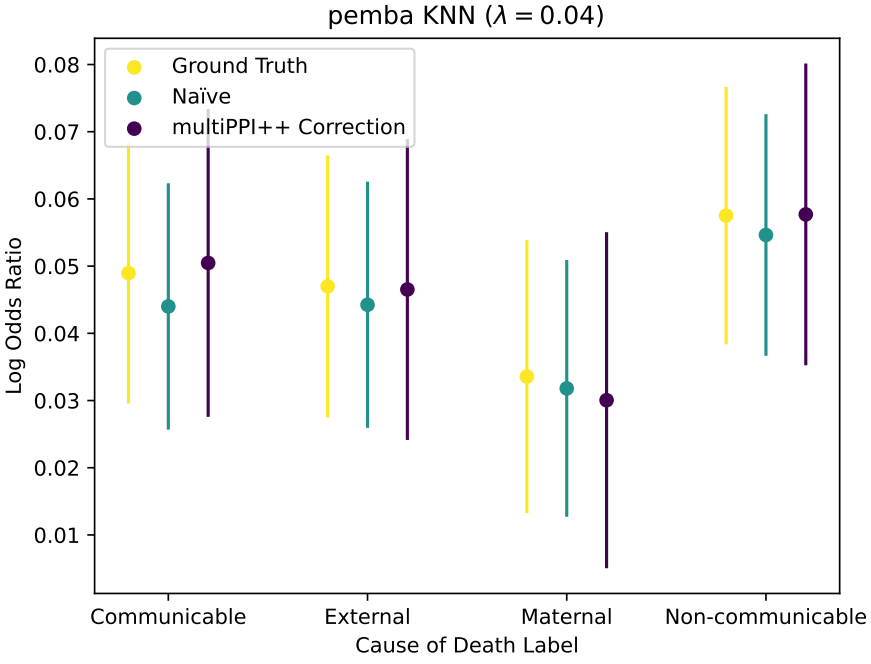


Figure C.34: Stratified 80/20 Split: Site/Model 22

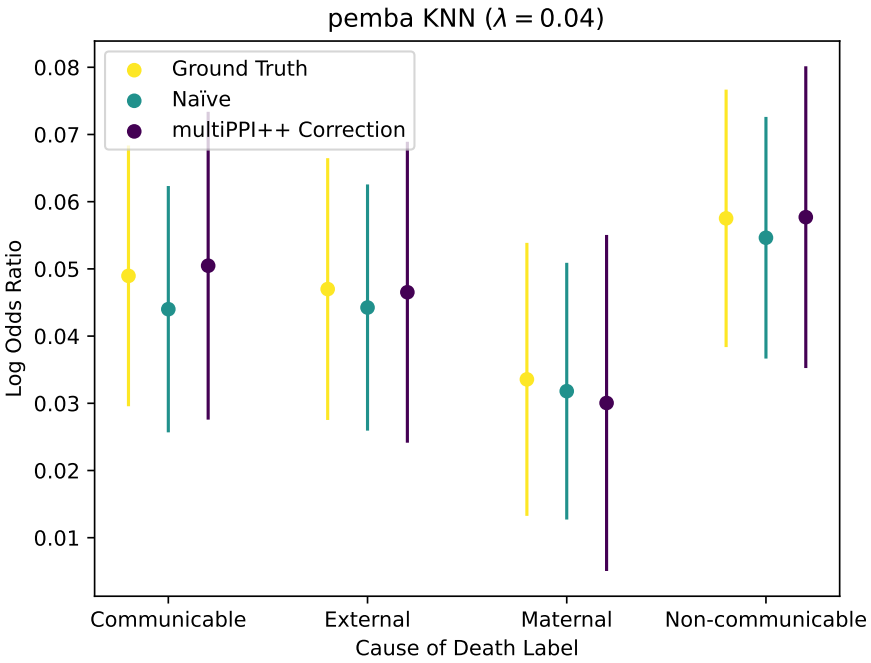


Figure C.35: Stratified 80/20 Split: Site/Model 22

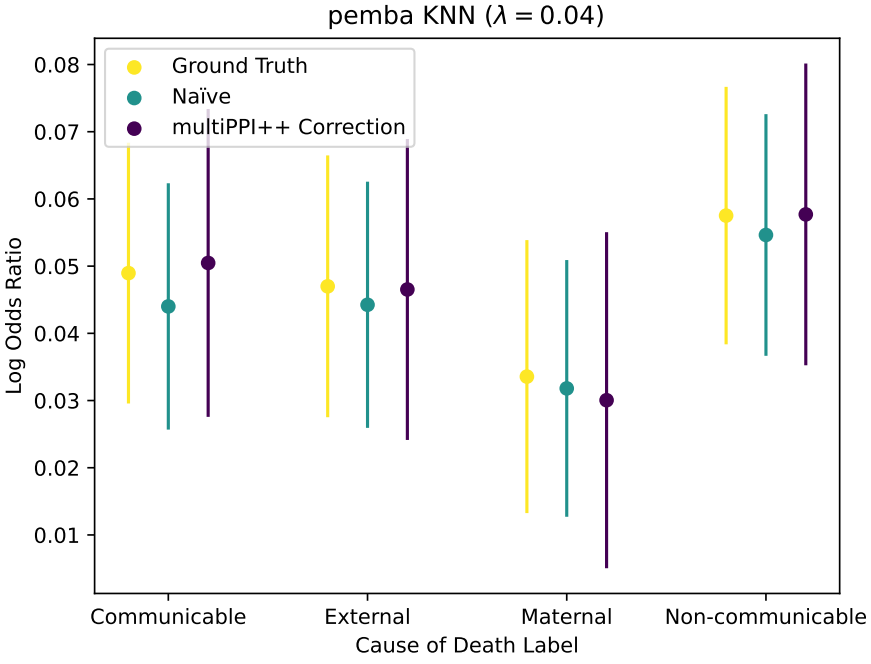


Figure C.36: Stratified 80/20 Split: Site/Model 22

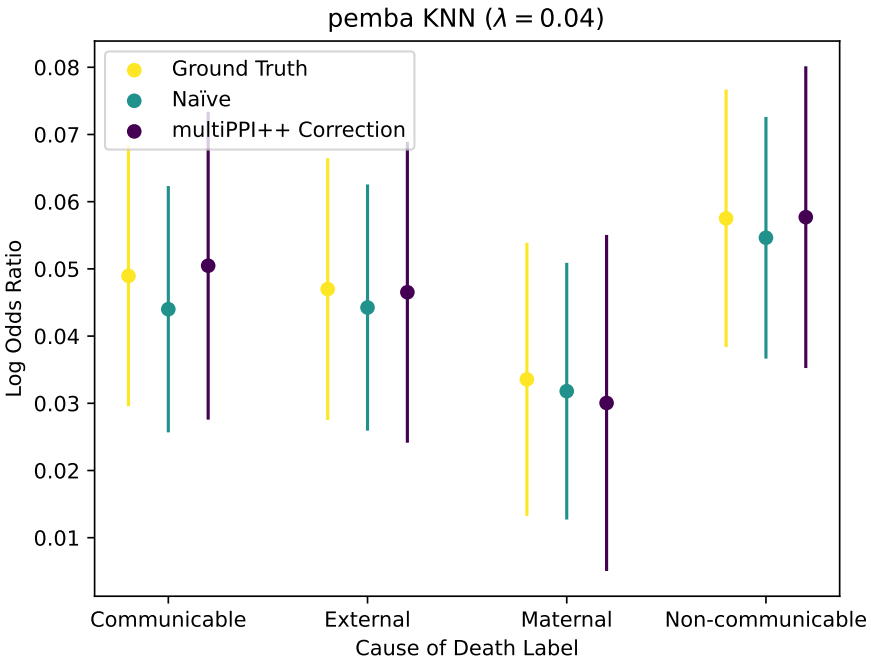


Figure C.37: Stratified 80/20 Split: Site/Model 22

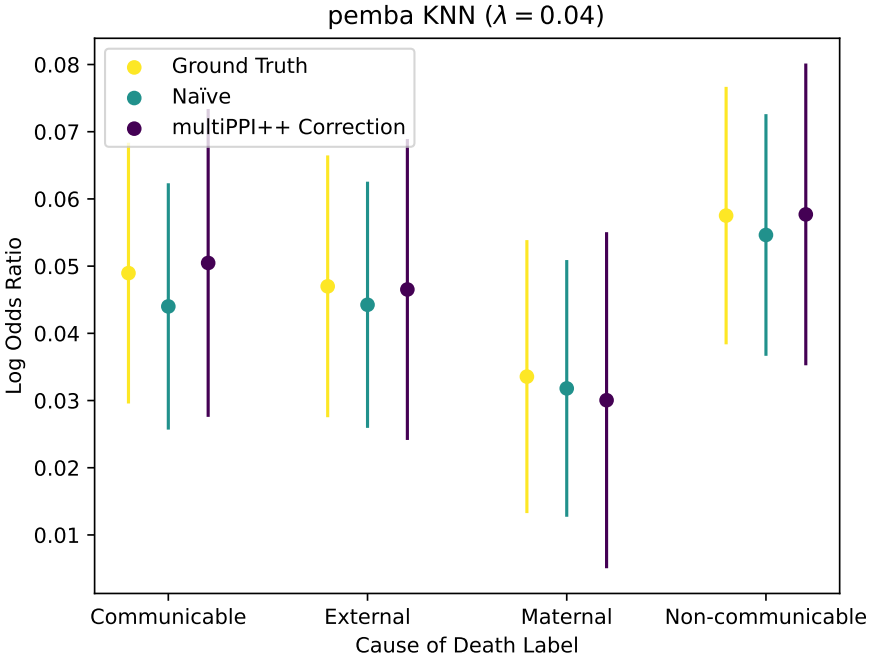


Figure C.38: Stratified 80/20 Split: Site/Model 22

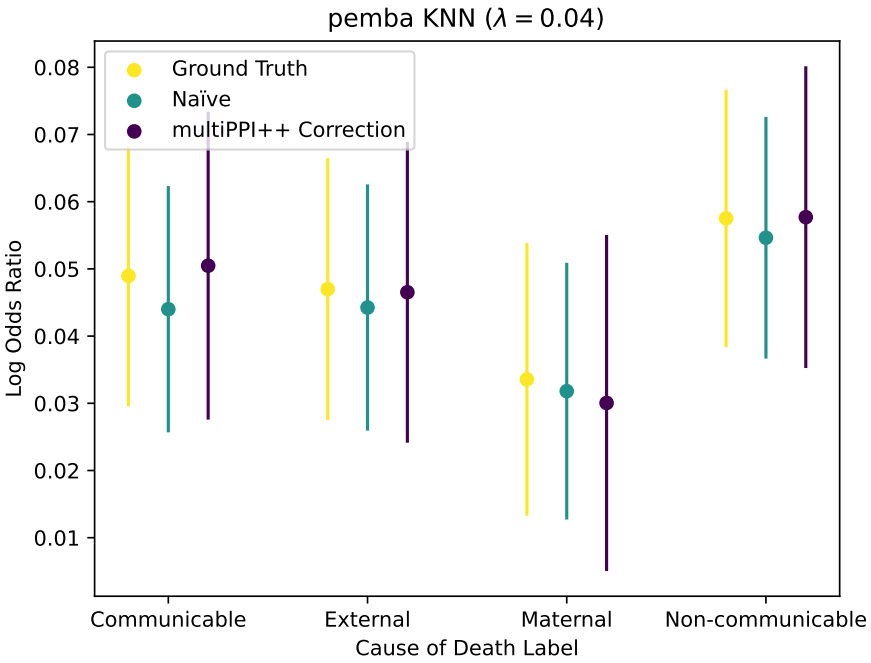


Figure C.39: Stratified 80/20 Split: Site/Model 22

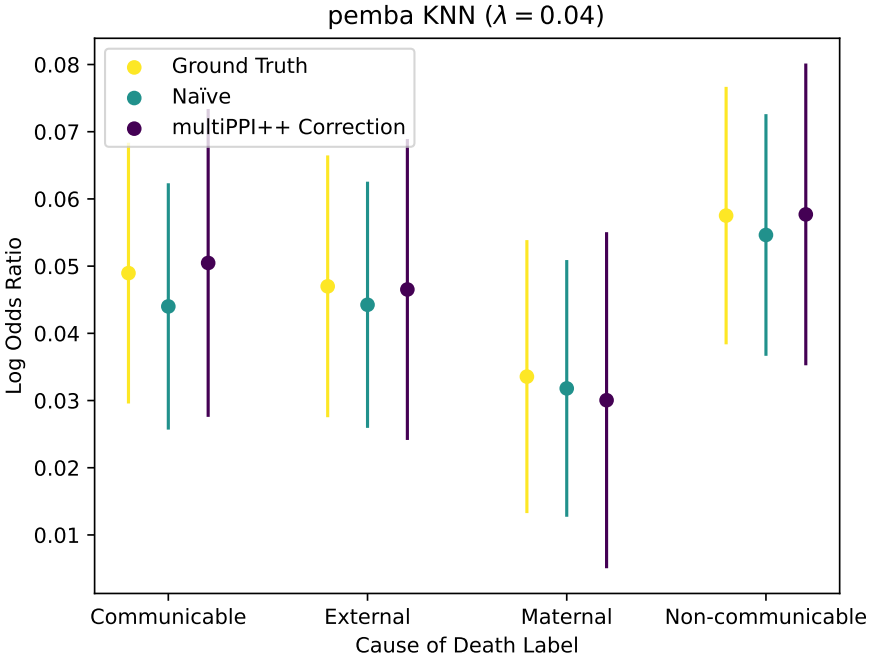


Figure C.40: Stratified 80/20 Split: Site/Model 22

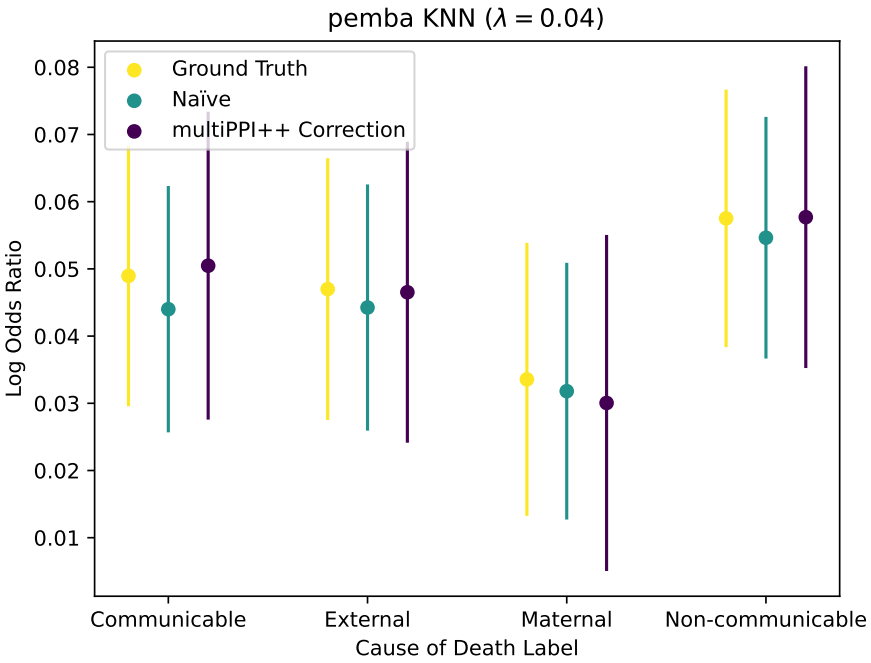


Figure C.41: Stratified 80/20 Split: Site/Model 22

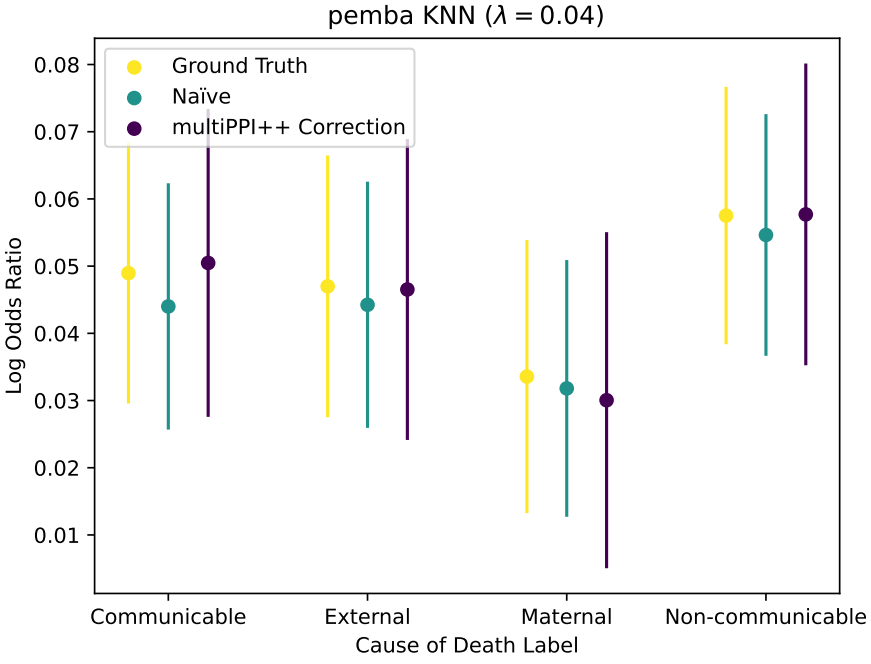


Figure C.42: Stratified 80/20 Split: Site/Model 22

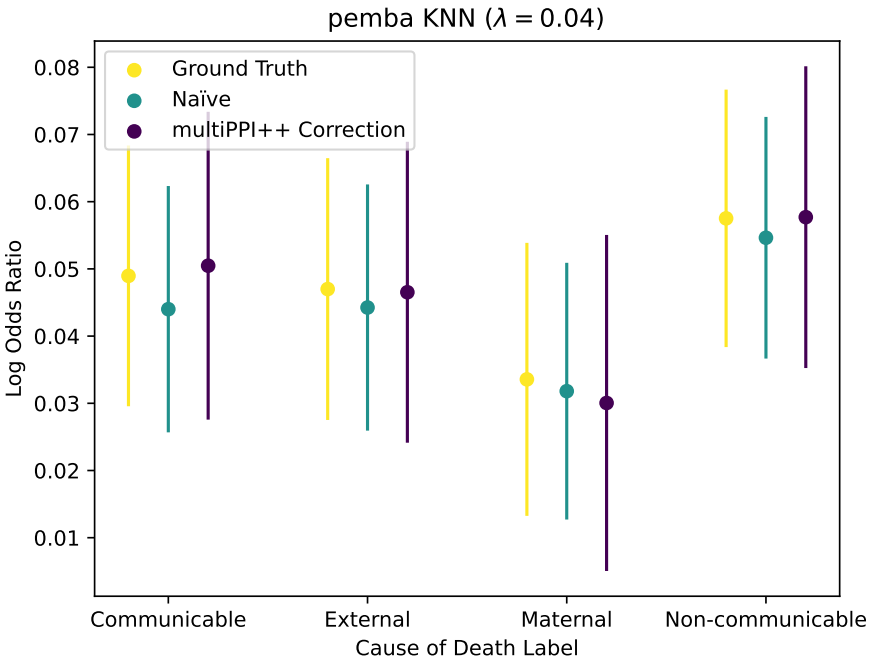


Figure C.43: Stratified 80/20 Split: Site/Model 22

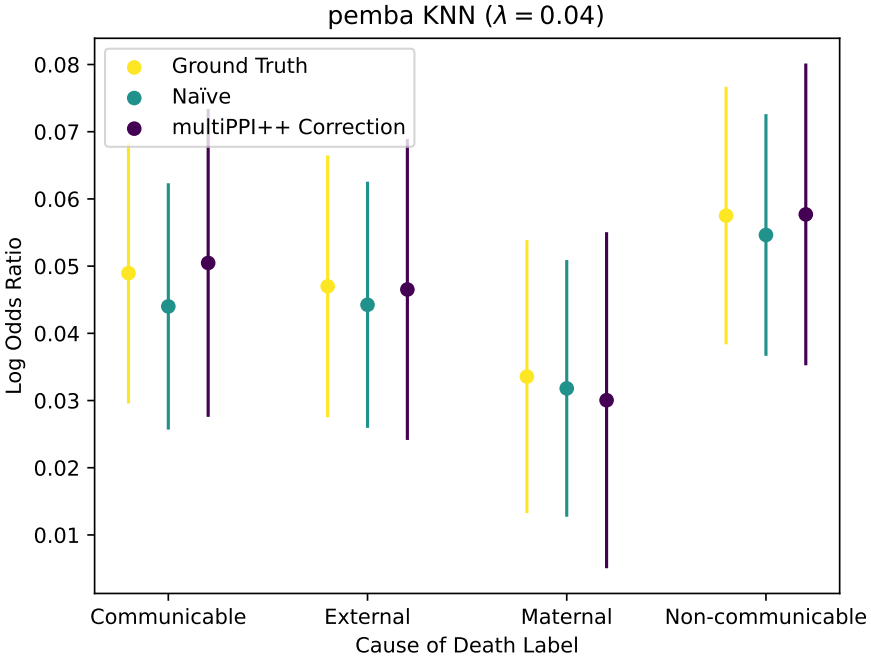


Figure C.44: Stratified 80/20 Split: Site/Model 22

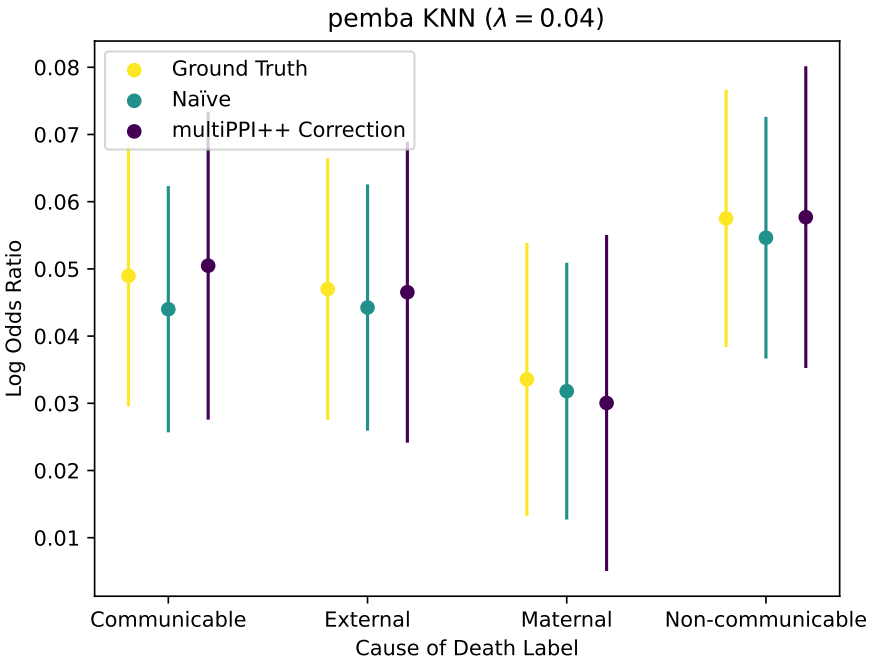


Figure C.45: Stratified 80/20 Split: Site/Model 22

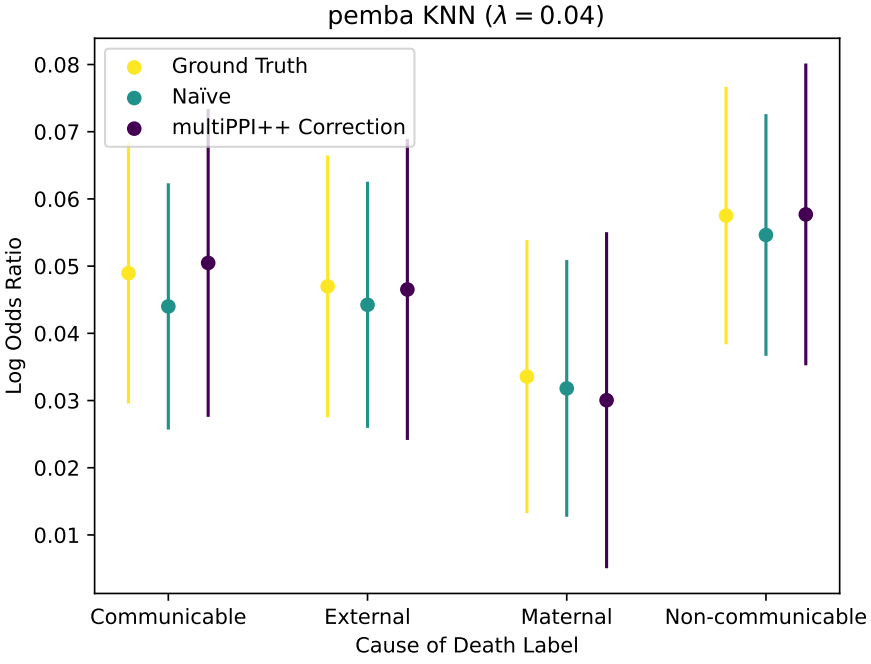


Figure C.46: Stratified 80/20 Split: Site/Model 22

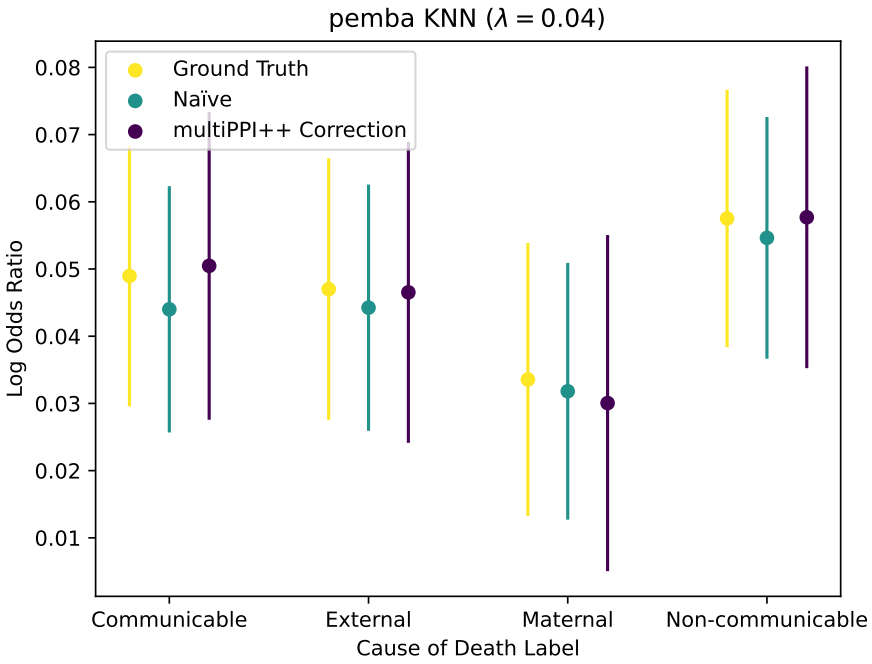


Figure C.47: Stratified 80/20 Split: Site/Model 22