# Handling onset age inconsistencies in longitudinal healthcare survey data

**Wanxin Li** W328LI@UWATERLOO.CA
*Department of Computer Science*
*University of British Columbia*
*Vancouver, British Columbia, Canada*

**Ming Yuan** MING.YUAN@STAT.UBC.CA
*Department of Statistics*
*University of British Columbia*
*Vancouver, British Columbia, Canada*

**Yongjin P. Park** YPP@STAT.UBC.CA
*Department of Statistics*
*University of British Columbia*
*Vancouver, British Columbia, Canada*

**Khanh Dao Duc** KDD@MATH.UBC.CA
*Department of Mathematics*
*University of British Columbia*
*Vancouver, British Columbia, Canada*

## Abstract

Longitudinal healthcare surveys frequently contain inconsistencies in self-reported onset ages, where participants report different ages for the same condition between enrollment and follow-up surveys. We propose two methods to handle this challenge. First, we introduce a procedure that aggregates inconsistency patterns to construct participant-level reliability scores, enabling researchers to stratify participants and prioritize analysis on high-reliability cohorts. Second, we present a Bayesian adjustment method that models enrollment and follow-up reports as noisy observations of a latent true onset age, producing adjusted estimates for the inconsistent observations that account for age-dependent and inter-survey-time effects. We evaluate both methods using data from the Canadian Partnership for Tomorrow's Health, where 57.1% of participants exhibit onset age inconsistencies. In general, both methods substantially strengthen correlations between biologically related conditions and improve predictive performance across classification and regression tasks. In addition, high-reliability cohorts from reliability score-based stratification reveal more coherent and interpretable disease clustering networks, and Bayesian adjustment shows particularly notable gains when multiple inconsistent variables are adjusted simultaneously. Finally, we provide guidance on choosing between these methods for healthcare practitioners.

## 1. Introduction

Longitudinal healthcare surveys are important for understanding disease etiology and developing predictive models for population health (Caruana et al., 2015; Grant et al., 1995;

Maddox and Douglass, 1973). These surveys collect self-reported information across survey waves, often spanning years or decades. A key variable in many of these surveys is the self-reported onset age, the age at which a participant first experienced or was diagnosed with a particular condition (Benjamin et al., 2017; Veenstra and Syse, 2012), which is important for life-course epidemiology and risk prediction (Skybova et al., 2022).

An onset age inconsistency occurs when a participant reports different onset ages for the same condition across survey waves. It is a type of measurement error as self-reported responses usually suffer from memory lapses, recall bias and careless responding (Shillington et al., 2010). For example, a participant might report developing diabetes at age 45 during enrollment but report age 52 during follow-up. Studies have shown that such inconsistencies are prevalent across large-scale cohorts (Johnson and Mott, 2001; Parra et al., 2003), yet they pose fundamental challenges: Discarding all inconsistent records can cause substantial data loss, while retaining them introduces measurement error that attenuates effect estimates. While previous work has quantified reliability patterns at the disease level (Johnson and Mott, 2001; Farrer et al., 1989), proposed deterministic reconciliation rules (Andreacchi et al., 2025), and developed structured interview techniques to improve prospective data collection (Axinn and Chardoul, 2021; Caspi et al., 1996), these approaches fail to quantify participant-level reliability or provide statistically principled adjustments that account for age-dependent measurement errors in existing datasets.

### Generalizable insights about machine learning in the context of healthcare

Longitudinal healthcare surveys are a common data source for machine learning-based health research; however, self-reported variables frequently contain measurement error that is heterogeneous across participants, conditions, and time, a challenge that is not addressed systematically. This work demonstrates that principled handling of such inconsistencies, rather than naive imputation or record discarding, yields meaningful gains in predictive performance and association discovery. Specifically, our findings suggest the following transferable lessons for healthcare practitioners working with longitudinal health data:

- **Participant-level data quality scores can serve as reusable pipeline inputs**: By aggregating inconsistency magnitudes across onset age variables, reliability scores can be computed once and used to stratify participants into high- and low-reliability cohorts across many downstream tasks, improving model performance without task-specific data cleaning.
- **Probabilistic noise modeling improves downstream predictive performance, scaling with multiple adjustments**: Our Bayesian adjustment method treats enrollment and follow-up reports as noisy observations of a latent true onset age, producing variance-weighted posterior estimates that consistently improve both classification and regression performance, with downstream errors decreasing progressively as multiple inconsistent variables are adjusted sequentially.
- **Method selection should be driven by data and task characteristics**: The two proposed methods are complementary, and we provide practical guidance for choosing between these methods based on sample size constraints, interpretability requirements, and variable characteristics.

## 2. Background and related works

### 2.1. Existing work for handling inconsistent observations in longitudinal healthcare data

Existing approaches for handling inconsistent longitudinal healthcare observations fall into three categories. First, reliability quantification studies measure the consistency of self-reported health information through test-retest assessments and validation against medical records (Johnson and Mott, 2001; Farrer et al., 1989; Louis et al., 2023; Raphael and Marbach, 1997; Dorn et al., 2013; Shillington et al., 2010). While these studies identify factors affecting reliability (e.g., disease type, recall period), they characterize reliability at the disease level and do not translate these patterns into participant-level metrics. Second, deterministic reconciliation applies rule-based adjudication to resolve contradictions, such as enforcing chronic-condition persistence or prioritizing specific data sources (Andreacchi et al., 2025). However, these approaches impose fixed solutions without quantifying uncertainty. Third, life history calendars use structured interview techniques to improve recall accuracy through temporal anchoring (Axinn and Chardoul, 2021; Caspi et al., 1996), but require prospective implementation and cannot be applied to existing data. In the broader statistical literature on longitudinal data, measurement error models treat repeated measures as noisy observations of a latent true value and estimate parameters via likelihood-based methods (Gelman et al., 1995; Quinio and Lam, 2021; Carroll and Stefanski, 1990). However, these models do not incorporate onset age-specific characteristics, such as age-dependent or inter-survey time effects. Moreover, they lack participant-level interpretability because they do not produce adjusted estimates for individual observations. Our work addresses these limitations by developing methods that quantify participant-level reliability and provide adjusted onset age estimates that account for age-dependent and inter-survey time effects.

### 2.2. CanPath

CanPath is a longitudinal healthcare survey study designed to investigate chronic disease prevention across Canada. The entire CanPath cohort contains over 330,000 participants. Our access request, based on variable availability and missingness criteria, yielded data for 97,408 participants. The CanPath study collects comprehensive data, including demographics, lifestyle factors (e.g., physical activity, smoking, alcohol consumption), socioeconomic variables (e.g., education, income), anthropometric measurements (e.g., body mass index), and self-reported disease histories. Participants completed baseline questionnaires at enrollment and were followed through subsequent surveys, typically 5-10 years later. Self-reported onset ages were collected for 55 variables at both enrollment and follow-up. Full names and descriptions of these 55 variables are provided in Appendix A.

## 3. Methods

### 3.1. Reliability score-based stratification

We construct and use participant-level reliability scores using the following steps. Let $\mathbf{X} \in \mathbb{R}^{n \times m}$ denote the dataset where $n$ is the number of participants and $m$ is the number of

variables, with $X_{ij}$ denoting participant $i$'s value for variable $j$. Let $\mathcal{A} \subseteq \{1, \ldots, m\}$ denote the set of onset age variables observed at both enrollment and follow-up, with $|\mathcal{A}| = p$. For $j \in \mathcal{A}$, let $X_{ij}^{(e)}$ and $X_{ij}^{(f)}$ denote participant $i$'s reported onset ages for variable $j$ at enrollment and follow-up, respectively.

**Data preparation** We construct the age difference matrix $\mathbf{D} \in \mathbb{R}^{n \times p}$ with entries:

$$D_{ij} = X_{ij}^{(f)} - X_{ij}^{(e)}, \quad j \in \mathcal{A} \tag{1}$$

where $D_{ij}$ is missing if either $X_{ij}^{(f)}$ or $X_{ij}^{(e)}$ is unavailable. We exclude participants with a missing rate exceeding a threshold $\rho$, yielding $n'$ participants for the following.

**Matrix completion** We assume that participant reliability depends only on the magnitude of onset age discrepancies, not their direction (i.e., whether ages are over- or under-reported at follow-up). Under this symmetric discrepancy assumption, we apply SoftImpute (Mazumder et al., 2010) to impute missing values in $\mathbf{D}$, yielding a completed matrix $\hat{\mathbf{D}} \in \mathbb{R}^{n' \times p}$, then compute element-wise absolute values $\tilde{\mathbf{D}} = |\hat{\mathbf{D}}|$.

**Score construction** We compute the raw reliability score for participant $i$ using: $r_i = \sum_{j=1}^{p} |\tilde{\mathbf{D}}_{ij}|$. Larger values of $r_i$ indicate greater deviations from consistent response patterns.

**Normalization** To obtain the final reliability scores, we apply quantile normalization (Bolstad et al., 2003) to transform the raw scores to a uniform distribution on $[0, 1]$, with inversion such that higher values indicate greater reliability.

**Stratification** The stratification converts the continuous reliability scores into actionable cohorts, enabling researchers isolate high-reliability participants for analysis. When the reliability score distribution exhibits clear multimodality, we can apply k-means clustering (Hartigan and Wong, 1979) to automatically identify thresholds. For unimodal distributions, we recommend examining the distribution's spread; if narrowly peaked, stratification may be unnecessary. Otherwise, researchers can use percentile-based thresholds (e.g., median) based on the sample size constraints.

### 3.2. Bayesian adjustment

Alternatively, we can directly adjust inconsistent onset age observations by modeling measurement error and computing posterior estimates of the latent values.

For $j \in \mathcal{A}$, let $X_{ij}^{*}$ denote participant $i$'s latent true value for variable $j$. Let $a_i^{(e)}$ and $a_i^{(f)}$ denote participant $i$'s ages at enrolment and follow-up, respectively.

**Measurement error model** We model the observed onset ages at enrollment ($X_{ij}^{(e)}$) and follow-up ($X_{ij}^{(f)}$) as noisy realizations of a latent true value ($X_{ij}^{*}$):

$$X_{ij}^{(e)} \mid X_{ij}^{*}, a_i^{(e)} \sim \mathcal{N}\left(X_{ij}^{*}, {\sigma_j^{(e)}}^2\right), X_{ij}^{(f)} \mid X_{ij}^{*}, a_i^{(e)}, \Delta_i \sim \mathcal{N}\left(X_{ij}^{*}, {\sigma_j^{(f)}}^2\right), \tag{2}$$

where $\Delta_i = a_i^{(f)} - a_i^{(e)}$, which is the time gap between follow-up and enrollment. We assume the measurement errors are independent conditional on $X_{ij}^{*}$ because enrollment and follow-up responses represent separate recall events occurring years apart, and no access to previous responses.

**Variance parameterization** $\Delta_i$ as independent predictors of measurement variance. To encode the assumption that recall accuracy decreases with age and deteriorates further at follow-up (Rhodes et al., 2019), we parameterize the variances as:[1]

$$\sigma_j^{(e)^2} = \sigma_j^2 e^{\alpha_{j0}+\alpha_{j1}a_i^e}, \sigma_j^{(f)^2} = \sigma_j^{(e)^2} e^{\delta_{j0}+\delta_{j1}\Delta_i},$$

where $\alpha_{j1}, \delta_{j0}, \delta_{j1} \geq 0$.[2] The constraint $\alpha_{j1} \geq 0$ ensures enrollment variance increases with age, while $\delta_{j0}, \delta_{j1} \geq 0$ ensures follow-up variance exceeds enrollment variance and increases with longer inter-survey intervals. This parameterization also relies on enrollment age $a_i^{(e)}$ and survey gap $\Delta_i$ are independent of each other.[3]

**Parameter estimation** To estimate the variance parameters $\{\sigma_j^2, \alpha_{j0}, \alpha_{j1}, \delta_{j0}, \delta_{j1}\}$ for each onset age variable $j$, we leverage the observed age differences $D_{ij}$. Let $n'_j \leq n'$ denote the number of participants with observed values for both $X_{ij}^{(e)}$ and $X_{ij}^{(f)}$. Since $X_{ij}^{(e)}$ and $X_{ij}^{(f)}$ are conditionally independent given $X_{ij}^*$ and both normally distributed (see Equation 2), their difference follows (Cramér, 1999):

$$D_{ij} \mid a_i^{(e)}, \Delta_i \sim \mathcal{N}\left(0, \sigma_j^{(e)^2} + \sigma_j^{(f)^2}\right). \tag{3}$$

We estimate the parameters by maximizing the log-likelihood:

$$\ell(\sigma_j^2, \alpha_{j0}, \alpha_{j1}, \delta_{j0}, \delta_{j1}) = \sum_{i=1}^{n'_j} \log Prob\left(D_{ij} \mid a_i^{(e)}, \Delta_i\right),$$

where the sum is over all participants with observed values for both $X_{ij}^{(e)}$ and $X_{ij}^{(f)}$.

**Posterior imputation** Assuming a diffuse Normal prior on $X_{ij}^*$,[4] the posterior distribution is:

$$X_{ij}^* \mid X_{ij}^{(e)}, X_{ij}^{(f)}, a_i^e, \Delta_i \sim \mathcal{N}(\mu_{ij}, \tau_{ij}^2),$$

where the posterior mean and variance are:

$$\mu_{ij} = \tau_{ij}^2 \left(\frac{X_{ij}^{(e)}}{\sigma_e^2} + \frac{X_{ij}^{(f)}}{\sigma_f^2}\right), \tau_{ij}^2 = \left(\frac{1}{\sigma_e^2} + \frac{1}{\sigma_f^2}\right)^{-1}.$$

The adjusted value $\hat{X}_{ij}^* = \mu_{ij}$ is a variance-weighted average that assigns greater weight to the enrollment observation as it has lower estimated variance. Derivations are provided in Appendix B.

1. Log-linear variance models are standard in heteroskedastic regression literature (Harvey, 1976).
2. We state $\delta_{j0}, \delta_{j1} \geq 0$ as identifying assumptions need to interpret the weights.
3. An alternative approach is to parameterize measurement variance as a function of the time elapsed since the true onset (i.e., $a_i^{(e)} - X_{ij}^*$). This formulation introduces a circular dependency in estimation: the measurement variance would depend directly on the latent variable $X_{ij}^*$ that we wish to infer. A poor initial estimate of $X_{ij}^*$ would distort the variance estimates, which in turn would reweight the observations and further bias the posterior mean, complicating parameter estimation and convergence.
4. This reflects the absence of strong condition-specific prior knowledge and ensures the posterior is driven by the two observed reports.

## 4. Synthetic experiments and results

### 4.1. Reliability score-based stratification

We validated the reliability score calculation procedure on synthetic data with ground-truth inconsistency levels. Across datasets ($n \in \{100, 200, 300, 400\}$, $p \in \{20, 30, 40\}$, informative missingness), we compared three variants: (1) our full method, (2) our method without SoftImpute, our method replacing Softimpute with (3) mean imputation (Little and Rubin, 2019), (4) kNN imputation (Batista, 2002), (5) Multivariate Imputation by Chained Equations (Azur et al., 2011), and (6) our method replacing simple averaging with variance-based weighting. Performance was measured by rank alignment across missingness levels $\{0.5, 0.6, 0.7, 0.8, 0.9\}$. Our study demonstrates that SoftImpute+avg is the best in the presence of missing data across all tasks on ranking alignment; it achieves an average score of 0.829, outperforming (3) and (4), while substantially surpassing (2), (4) and (6).

Our method (SoftImpute + simple averaging) achieved consistent rank alignment of 0.78–0.82, substantially outperforming the no-imputation variant (0.22–0.41), which degrades sharply with increasing missingness. Variance-based weighting performed nearly identically (0.77–0.81) without added benefit. See Appendix C for details. These results confirm our choice of pipeline components to uncover the ground-truth inconsistencies.

### 4.2. Bayesian adjustment

We validated the Bayesian adjustment method on synthetic data generated under the exact measurement error model assumptions specified in Section 3.2, across seven scenarios varying sample size ($N \in \{2{,}000, 10{,}000\}$), noise level ($\sigma_0 \in \{0.5, 1.0, 2.0\}$), age dependence ($\alpha_1 \in \{0.010, 0.050\}$), and inter-survey gap length (gap range $\in \{[2, 20], [10, 30]\}$ years). True latent onset ages were drawn from $\mathcal{N}(35, 10)$, and enrollment and follow-up measurements were simulated as noisy observations using Equation 2. Across all scenarios, Bayesian adjustment matched or outperformed enrollment-only, follow-up-only, and simple-average baselines, representing an average MAE improvement of 1.38% to 34.07% relative to the best and worst performing baseline, respectively. See Appendix D for details. These results show that, under the assumed model, Bayesian adjustment consistently recovers latent true onset ages more accurately than any single observation or their naive average.

## 5. Empirical experiments, assumption checking and implementation

### 5.1. Reliability score-based stratification

**Experiments** We evaluated reliability score-based stratification on two downstream tasks: association discovery and predictive modeling. For each task, we stratified participants into high- and low-reliability cohorts based on the median reliability score, excluding participants with missing values on the task-specific variables. For all onset ages mentioned in this section, they refer to the onset ages during enrollment.[5]

In Section 6.2.2, for association discovery, within each of the high- and low-reliability cohorts, we examined pairwise Pearson correlations between onset ages of biologically related

5. This is different from Section 5.2 where we used both the onset ages during enrollment and follow-up.

conditions and constructed correlation networks to identify disease clusters. The correlation analyses include:

($co_{11}$) Asthma, high cholesterol, high blood pressure, heart attack, stroke onset ages (Garg et al., 2023; Tattersall et al., 2016; Wang et al., 2020; Khoja et al., 2023; Domanski et al., 2020) (mean 458 participants per pairwise analysis);

($co_{12}$) Hearing loss, swimmer's ear, tinnitus and vertigo onset ages (Wiegand et al., 2019; Azevedo et al., 2022; Ihler et al., 2024; Reisinger et al., 2023; Miura et al., 2017) (mean 39 participants per pairwise analysis).

For network analysis:

($co_{13}$) Correlation networks on 45 disease onset age variables with positive correlations out of 55 onset age variables (mean 151 participants per pairwise analysis).

In Section 6.2.3, for predictive modeling, for each of the high- and low-reliability cohorts, we evaluated classification tasks (predicting disease status) and regression tasks (predicting onset age) using 5-fold cross-validation. For classification, we applied Logistic Regression, Random Forest (RF), Decision Tree (DT), Gradient Boosting (GB), and Support Vector Machine (SVM) with precision and recall as metrics. For regression, we applied Linear Regression, Lasso, Ridge, RF, and SVM with MAE and Root Mean Squared Error (RMSE) as metrics. The classification tasks include:

($cl_{11}$) Predicting depression status using sex, age, and anxiety status (2198 participants);

($cl_{12}$) Predicting high blood sugar status using age, sex, high cholesterol status, diabetes status, income, physical activity level, smoking frequency, and alcohol frequency (2953 participants);

($cl_{13}$) Predicting diabetes status using high blood sugar onset age, and high blood pressure onset age (498 participants).

The regression tasks include:

($re_{11}$) Predicting depression onset age using age, sex, and anxiety onset age (500 participants);

($re_{12}$) Predicting high blood sugar onset age using age, sex, high cholesterol onset age, physical activity level (493 participants);

($re_{13}$) Predicting hearing loss onset age using tinnitus onset age, and tinnitus treatment status (1032 participants).

**Assumption checking** To verify the symmetric discrepancy assumption in matrix completion step, we conducted Wilcoxon signed-rank tests (Woolson, 2007) for the 55 onset age with paired observations and reported Cohen's $d$ (Cohen, 2013) as a separate effect-size criterion to guard against large-sample inflation of statistical significance (negligible $|d| < 0.2$, small $0.2 \leq |d| \leq 0.5$, medium $0.5 \leq |d| \leq 0.8$, large $|d| \geq 0.8$ ). Across all 55 variables, obsessive-compulsive disorder has $|d| = 0.84$ and is statistically significant, making it one

condition with meaningful directional bias; schizophrenia has fewer than 5 non-zero paired observations and cannot be submitted to the Wilcoxon test. No other variables simultaneously reach significance and $|d| \geq 0.2$. Hence, in general, the symmetric discrepancy assumption holds for our dataset.

**Implementation** In data preparation, we set the missing rate threshold $\rho = 53/55$, ensuring every retained participant has at least two observed onset age differences and thus contributes meaningful signal to the matrix completion step, leaving 25,018 participants for analysis. [6] The implementation and experiment code for both methods are at https://github.com/wxli0/canpath.git.

### 5.2. Bayesian adjustment

**Experiments** We evaluated Bayesian adjustment on association discovery and predictive modeling tasks. In Section 6.3.1, for association discovery, we computed Pearson correlations between pairs of onset age variables using enrollment variables, follow-up variables, and Bayesian-adjusted values (applied to both variables). We selected onset age variable pairs with established positive biological associations:

($co_{21}$) Anxiety and depression (Jacobson and Newman, 2017) (260 participants),

($co_{22}$) High blood pressure and heart attack (Psaty et al., 2001) (474 participants),

($co_{23}$) High cholesterol and high blood pressure (Grundy et al., 2005) (1554 participants).

In Section 6.3.2, for predictive modeling, we evaluated classification and regression tasks with and without Bayesian adjustment, and with regression calibration (RC) using 5-fold cross-validation. For each task, we used the set of models and metrics as in Section 5.1. In addition, we accompanied all metrics with 95% confidence intervals (CIs) computed via non-parametric bootstrapping across the cross-validation folds. Onset age variables with potentially inconsistent enrollment and follow-up observations are marked with $*$; these are the variables to which we applied Bayesian adjustment. The classification tasks include:

($cl_{21}$) Predicting depression status using age, anxiety onset age ($*$) (335 participants),

($cl_{22}$) Predicting diabetes status using age, high blood sugar onset age, and high blood pressure onset age ($*$) (579 participants).

The regression tasks include:

($re_{21}$) Predicting diabetes onset age using age, sex, high blood pressure onset age ($*$), and high cholesterol onset age ($*$) (314 participants),

($re_{22}$) Predicting high cholesterol onset age using age, sex, high blood pressure onset age ($*$), and high blood sugar status (1380 participants).

6. In Appendix E, we assessed Social Determinants of Health variables distribution between included and excluded partcipants, and confirmed that this exclusion does not deliberately introduce systematic demographic bias and impedes the generalizability of our method.

The tasks, variables, and sample sizes differ between Section 5.1 and here. We clarify that the two methods are not intended to be directly compared against each other, see discussion in Section 7. The difference in tasks, variables, and sample sizes across the two methods also arises naturally from their distinct data requirements. Reliability score-based stratification requires participants to have sufficient non-missing onset age observations across variables to construct a meaningful reliability score, while Bayesian adjustment requires participants to have observed values at both enrollment and follow-up for the specific variables being adjusted. Applying the missingness filters appropriate to each method necessarily yields different participant subsets. When fewer than 20 participants remained for certain variable combinations, we selected alternative tasks and variables.

**Assumption checking** We verified three assumptions underlying the Bayesian adjustment model. (1) Gaussian structure: One-sample $t$-tests on age differences $D_{ij}$ across all four tasks confirmed zero mean ($p > 0.05$). The excess kurtosis (5.1–14.4) is consistent with known mixture-of-Gaussians heavy tails (Rocha et al., 2012). (2) Independence of enrollment age and survey gap: Variance Inflation Factors $\approx 1.0$ ($R^2 \approx 0.0$–$0.04$) across all tasks, indicating negligible collinearity between $a_i^{(e)}$ and $\Delta_i$. (3) Exponential variance structure: Log-linear regression of residual variance against $a_i^{(e)}$ confirmed exponential growth in all four tasks ($R^2 = 0.68$–$0.93$, $p < 0.001$). Because 92.5% of participants have $\Delta_i \in [4, 7]$ years with a standard deviation of 1.1, the inter-survey interval component $\delta_{j1}$ is not identifiable. Indeed, across all experiments, $\delta_{j1}$ is statistically indistinguishable from zero (all $p > 0.05$). In this case, our model reduces to a simpler but still principled form with age-dependent enrollment variance. (4) Identifying assumption: On every onset-age variable in Section 5.2, we verified that $Var(X^{(e)}) < Var(X^{(f)})$ on all paired variables. In addition, refitting with $\delta_{j0}, \delta_{j1}$ unconstrained on the four tasks in Section 5.2 changes the weights substantially but leaves the posterior mean $\hat{X}_{ij}^*$ essentially unchanged (Pearson, Spearman $\geq 0.95$; mean bias $\leq 0.20$ yr).

**Implementation** In parameter estimation, to enforce the constraints $\alpha_{j1}, \delta_{j0}, \delta_{j1} \geq 0$ during maximum likelihood estimation, we reparameterized these parameters using exponential transformations. We maximized the log-likelihood using multiple solvers (L-BFGS-B (Byrd et al., 1995), SLSQP (Kraft, 1988), and TNC (Nash, 1984)) and selected the solution with the lowest negative log-likelihood.

## 6. Empirical results

### 6.1. Quantification of onset age inconsistencies in CanPath data

We quantified onset age inconsistencies in CanPath data by computing the difference between self-reported onset ages at enrollment and follow-up for each condition. Figure 1 shows the distribution of these differences across 55 onset age variables in CanPath data, ordered by variance. A non-zero difference indicates an inconsistency, with larger absolute values suggesting greater inconsistency. Variance differs substantially across conditions, with the age differences for some conditions (e.g., arthritis, breast cancer) showing notably higher variance than others.

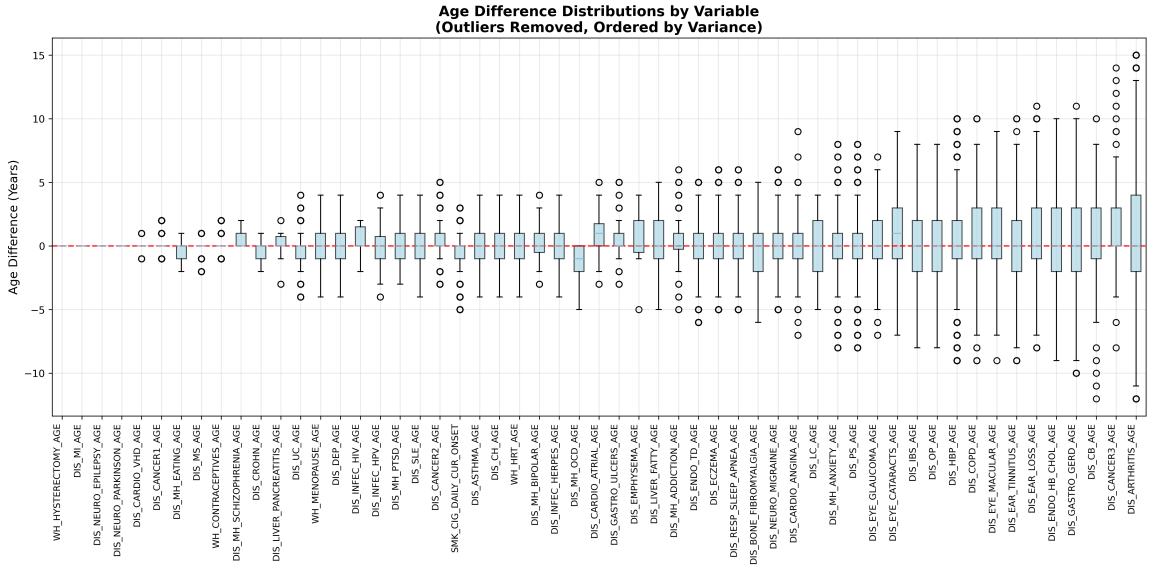


Figure 1: Box plots of enrollment and follow-up age difference distributions for onset age variables in CanPath data, with outliers removed and ordered by variance. Descriptions of the onset age variables are provided in Appendix A.

### 6.2. Reliability score-based stratification

#### 6.2.1. Uniform reliability score distribution

Appendix F shows the distribution of original reliability scores without quantile normalization where we do not observe extreme outliers. Figure 2(a) shows the distribution of reliability scores across all participants after normalization. The distribution is approximately uniform over [0,1] with mean 0.501 and median 0.499. This uniform distribution enables effective stratification of participants for our later experiments, as the absence of peaks or clustering at particular values ensures balanced cohort sizes for any subset of participants in task-specific evaluations.

#### 6.2.2. Stronger correlations and more coherent disease clustering in high-reliability cohorts

Figures 2(b) and (c) show heatmaps where each cell shows the difference in correlation (high-reliability cohort minus low-reliability cohort) for pairs of onset ages in (b) chronic diseases ($co_{11}$) and (c) hearing conditions ($co_{12}$), respectively, where positive values indicate stronger correlations in the high-reliability cohort. Across both disease groups, almost all correlations are consistently stronger in the high-reliability cohort, with differences ranging from 0.01 to 0.95. This indicates that the high-reliability cohorts using reliability score-based stratification produce stronger associations consistent with known biological relationships.

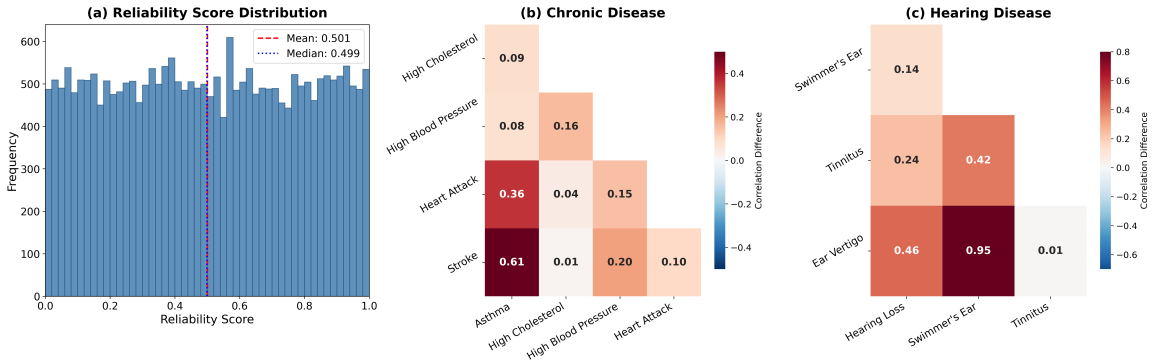


Figure 2: Reliability score distribution and onset age correlation differences between cohorts. (a) Distribution of reliability scores used to split participants into high- and low-reliability cohorts at the median. (b-c) Difference in correlations (high-reliability minus low-reliability) for pairs of onset ages in (b) chronic diseases and (c) hearing conditions. Positive values (red) indicate stronger correlations in the high-reliability cohort.

Figure 3 presents disease onset age correlation networks for each cohort ($co_{13}$), with communities identified using the Louvain algorithm (Blondel et al., 2008) with resolution parameter set to 1. To ensure that the observed clustering differences are not artifacts of the chosen community detection hyperparameters, we conducted a sensitivity analysis across varying resolution parameters in Appendix G. Across all configurations, the high-reliability network consistently demonstrates superior cluster quality and biological coherence. We classified diseases into 12 categories based on the medical system (e.g., Cardiovascular, Respiratory, Gastrointestinal/Hepatic, Mental Health) using keyword matching on variable names. Full details are provided in Appendix H. While both cohorts yield five clusters across 45 onset age variables, the high-reliability network exhibits markedly greater biological coherence[7]: The average proportion of diseases belonging to the dominant disease category within each cluster increases from 30.9% to 43.8%, while cluster entropy decreases from 2.23 to 1.86,[8] indicating that diseases are more consistently grouped with biologically related conditions in the high-reliability cohort.

The improved clustering is particularly evident when examining the variables in specific clusters. In cluster 3 of the high-reliability network, gastrointestinal conditions (Crohn's disease (CROHN), ulcerative colitis (UC), irritable bowel syndrome (IBS), persistent acid reflux (GASTRA_GERD), stomach (or duodenal) ulcers (GASTRA_ULCERS), and pancreatitis (LIVER_PANCREATITIS)) cluster together (60% gastrointestinal), whereas in the low-reliability network these conditions are scattered across multiple clusters. Similarly, in cluster 2 of the high-reliability network, cardiovascular and metabolic conditions (high blood pressure (HBP), high cholesterol (ENDO_HB_CHOL), heart attack (MI), angina

7. Full definitions of the coherence metrics are provided in Appendix I.
8. The detailed cluster composition is provided in Appendix J.

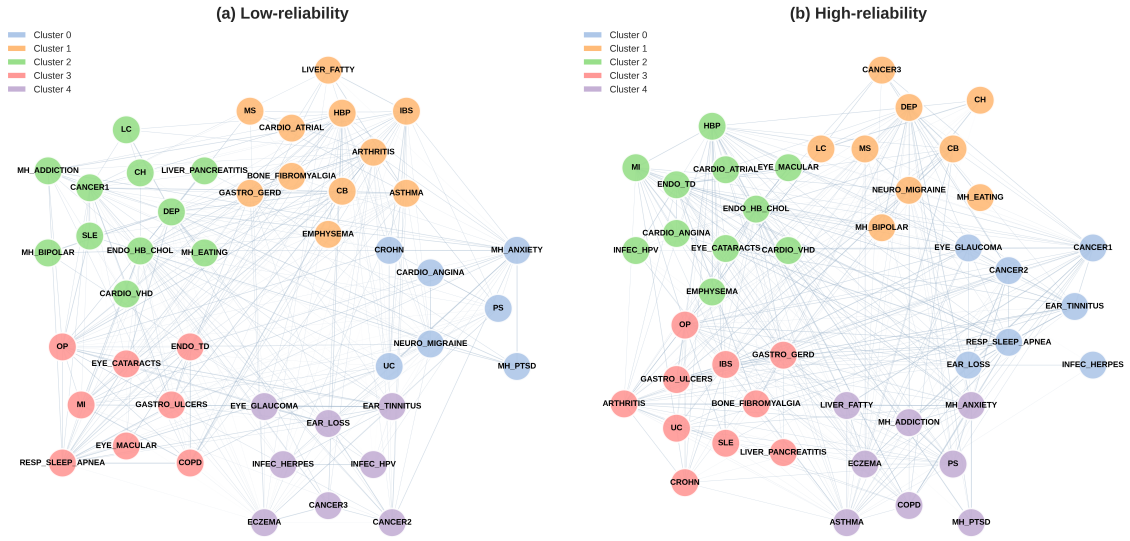


Figure 3: Disease onset age correlation networks stratified by reliability scores. Cluster networks showing communities of diseases with correlated ages of onset for (a) low-reliability and (b) high-reliability cohorts. Each node represents a disease, and edges indicate positive correlations between onset ages. Nodes are colored by cluster membership identified using the Louvain community detection algorithm. The average proportion of diseases in the dominant category within each cluster is 30.9% (low-reliability) and 43.8% (high-reliability), with cluster entropy of 2.23 (low-reliability) and 1.86 (high-reliability).

(CARDIO_ANGINA), atrial fibrillation (CARDIO_ATRIAL)) form a cohesive cluster in the high-reliability network but are more diffusely distributed in the low-reliability network. These findings demonstrate that the high-reliability cohort significantly strengthens pairwise correlations and reveals clearer disease community structure.

### 6.2.3. Improved predictive performance in high-reliability cohorts

Table 1 summarizes cross-validation performance for classification ($cl_{11}$–$cl_{13}$) and regression ($re_{11}$–$re_{13}$) tasks across cohorts. For each task, we selected the best model by F1 score (classification) or RMSE (regression), with model selection performed independently for each cohort; tie-breaking rules are available in Appendix K. To ensure these differences reflect data quality rather than demographic confounds, we confirmed in Appendix L that SDoH differences between high- and low-reliability cohorts are negligible to small across all tasks. Full predictive results are provided in Appendix M.

For regression tasks, the high-reliability cohort consistently yields substantially lower prediction errors, with MAE improvements ranging from 0.90 – 1.75 years and RMSE improvements from 0.71 – 1.82 years. Classification results are more mixed. For high blood

Table 1: Cross-validation predictive performance on low- versus high-reliability cohorts for classification and regression tasks. Cohorts are divided by median reliability score. The best model is selected based on F1 score (classification) or RMSE (regression) on high- and low-reliability cohorts separately. High-reliability cohort rows are highlighted in gray.

| ID | Target | Best model | Cohort | Metrics | |
|---|---|---|---|---|---|
| | | | | Precision | Recall |
| $cl_{11}$ | Depression | Logistic | Low | 0.745 | 0.731 |
| | | Logistic | High | 0.733 | 0.710 |
| $cl_{12}$ | High blood sugar | Logistic/SVM | Low | 0.928 | 0.928 |
| | | Logistic/SVM | High | 0.957 | 0.957 |
| $cl_{13}$ | Diabetes | RF | Low | 0.681 | 0.719 |
| | | GB | High | 0.751 | 0.759 |

| ID | Target | Best model | Cohort | Metrics | |
|---|---|---|---|---|---|
| | | | | MAE | RMSE |
| $re_{11}$ | Depression | Linear | Low | 7.052 | 9.646 |
| | | Linear | High | 5.306 | 7.825 |
| $re_{12}$ | High blood sugar | Ridge | Low | 6.560 | 8.782 |
| | | Ridge | High | 5.656 | 8.077 |
| $re_{13}$ | Hearing loss | Linear | Low | 7.847 | 11.477 |
| | | Linear | High | 6.841 | 10.610 |

sugar and diabetes prediction, the high-reliability cohort shows clear improvements in both precision and recall. However, for depression prediction, the low-reliability cohort slightly outperforms the high-reliability cohort (0.745 vs. 0.733 in precision, and 0.731 vs. 0.710 in recall). We attribute this to systematic underrepresentation of individuals with depression in the high-reliability cohort. Of the 55 onset age variables used to construct reliability scores, only 7 are mental health-related. If mental health variables exhibit a different inconsistency pattern, the reliability score is dominated by non-mental-health variables, causing individuals with depression to be disproportionately assigned low reliability scores. To verify this, we tested the association between disease status and reliability scores across all disease variables using Cohen's $d$ (negative values indicate lower reliability scores among those with the disease). Depression is indeed associated with lower reliability scores (mean 0.446 vs. 0.523 for those without; $d = -0.27$). Moreover, mental health conditions show systematically larger negative effects than physical conditions (anxiety $d = -0.40$; addiction $d = -1.01$; vs. asthma $d = -0.06$; eczema $d = -0.10$; thyroid disease $d = -0.21$), suggesting that response variability in mental health self-reporting carries a meaningful clinical signal rather than reflecting noise.

### 6.3. Bayesian adjustment

#### 6.3.1. Strengthened correlations between biologically related conditions

Table 2 shows correlations computed using enrollment variables, follow-up variables, Bayesian-adjusted values and RC. Across all pairs, the correlation for Bayesian-adjusted values exceeds both the enrollment variables correlation and the follow-up variables. This indicates our Bayesian adjustment method reveals more stable patterns consistent with underlying disease relationships.

Table 2: Pearson correlation coefficients between onset ages of variable 1 and 2 using enrollment variables, follow-up variables, and Bayesian-adjusted values.

| ID | Variable 1 | Variable 2 | Enrollment | Follow-up | Bayesian adjustment |
|---|---|---|---|---|---|
| $co_{21}$ | Anxiety | Depression | 0.670 | 0.645 | 0.692 |
| $co_{22}$ | High blood pressure | Heart attack | 0.507 | 0.518 | 0.536 |
| $co_{23}$ | High cholesterol | High blood pressure | 0.617 | 0.648 | 0.650 |

#### 6.3.2. Consistent improvements and ablation of multiple variable adjustment

Table 3 shows cross-validation performance for classification and regression tasks. We selected the best model by F1 score (classification) or RMSE (regression) independently for each setting: two unadjusted baselines (enrollment only and follow-up only) and Bayesian adjustment; tie-breaking rules are provided in Appendix K. The full results are provided in Appendix N.

Across all tasks, Bayesian adjustment consistently improves predictive performance. To investigate the impact of adjusting multiple inconsistent predictors, we conducted a single-task ablation study on diabetes onset age prediction ($re_{21}$) by sequentially adjusting $0 \to 1 \to 2$ variables (high blood pressure and high cholesterol onset ages). As the number of adjusted predictors increases, we observe a monotonic decrease in prediction error: MAE drops from 6.802 to 5.630, and then to 4.753, while RMSE decreases from 9.183 to 8.244, and then to 6.471. This indicates a clear benefit to adjusting multiple variables in this specific setting.

## 7. Conclusion and discussion

In this paper, we proposed two methods for handling onset age inconsistencies in longitudinal healthcare survey data. First, we constructed participant-level reliability scores by aggregating inconsistency magnitudes across onset age variables, enabling stratification of participants into high- and low-reliability cohorts. Second, we proposed a Bayesian adjustment method that models enrollment and follow-up onset ages as noisy measurements of a latent true onset age, producing variance-weighted posterior estimates for inconsistent observations. On synthetic data, reliability score-based stratification successfully uncovered

Table 3: Cross-validation predictive performance across three settings: unadjusted using enrollment data only, unadjusted using follow-up data only, and Bayesian-adjusted. The best model for each setting is selected by F1 score (classification) or RMSE (regression). 95% CIs are shown in brackets. The unadjusted results using follow-up data only are highlighted in light gray; Bayesian-adjusted results are highlighted in dark gray.

| ID | Target | Adjustment | Best model | Metrics | |
|---|---|---|---|---|---|
| | | | | Precision | Recall |
| $cl_{21}$ | Depression | No (enrollment) | SVM | 0.699 [0.619, 0.715] | 0.696 [0.621, 0.713] |
| | | No (follow-up) | Logistic | 0.695 [0.647, 0.709] | 0.693 [0.648, 0.704] |
| | | RC | SVM | 0.696 | 0.693 |
| | | Yes | SVM | 0.745 [0.654, 0.783] | 0.746 [0.657, 0.776] |
| $cl_{22}$ | Diabetes | No (enrollment) | GB | 0.658 [0.580, 0.711] | 0.699 [0.622, 0.734] |
| | | No (follow-up) | GB | 0.688 [0.591, 0.724] | 0.705 [0.634, 0.746] |
| | | RC | RF | 0.650 | 0.681 |
| | | Yes | RF | 0.685 [0.568, 0.706] | 0.713 [0.626, 0.730] |

| ID | Target | Adjustment | Best model | Metrics | |
|---|---|---|---|---|---|
| | | | | MAE | RMSE |
| $re_{21}$ | Diabetes | No (enrollment) | Linear | 5.566 [5.297, 5.979] | 7.580 [7.321, 8.122] |
| | | No (follow-up) | Ridge | 5.578 [5.363, 5.985] | 7.581 [7.340, 8.068] |
| | | RC | Ridge | 6.587 | 8.989 |
| | | Yes | Lasso | 4.625 [4.554, 4.952] | 6.471 [6.296, 6.975] |
| $re_{22}$ | High cholesterol | No (enrollment) | Linear/ Ridge | 5.157 [5.073, 5.273] | 6.901 [6.857, 7.002] |
| | | No (follow-up) | Linear/ Ridge | 5.312 [5.242, 5.420] | 6.985 [6.946, 7.078] |
| | | RC | Linear/ Ridge | 5.151 | 6.865 |
| | | Yes | Linear | 4.946 [4.892, 5.217] | 6.357 [6.277, 6.606] |

ground-truth inconsistency patterns across varying missingness levels, and Bayesian adjustment consistently matched or outperformed all baselines in recovering latent true onset ages. On CanPath data, high-reliability cohorts demonstrate stronger positive correlations among biologically related variables, more coherent and interpretable disease communities in clustering networks, and better predictive performance across classification and regression tasks, and Bayesian adjustment increases correlation estimates for biologically associated onset-age pairs and improves downstream predictive performance for both tasks.

These two methods address different needs for healthcare practitioners and are best viewed as complementary. Reliability score-based stratification is more appropriate when (1) the dataset is sufficiently large that deprioritizing or excluding participants in low-reliability cohorts does not impact the learning of tasks; we recommend that retained participants have observed onset age differences for at least two variables; (2) missingness across onset age variables is moderate, so that matrix completion can recover meaningful signal

(we recommend at least two); and (3) ease of deployment is the priority, since the scores can be computed once and reused across analyses. In contrast, Bayesian adjustment is more appropriate when (1) the sample size of participants with paired observations is sufficient for parameter estimation; (2) the researchers want to propagate uncertainty from the data into subsequent inference; and (3) tasks involve mental-health variables, where the response variability may be different from other variables, making variable-by-variable adjustment preferable to participant-level exclusion.

Two caveats need attention. Reliability score-based stratification may introduce selection bias by systematically excluding certain subpopulations. One should compare SDoH distribution between cohorts before excluding low-reliability participants, and carefully consider the equity implications of doing so; for example, whether a SDoH difference reflects a deliberate cohort imbalance, as illustrated in Section 6.2.3. Furthermore, as discussed in Section 6.2.3, one should be cautious when interpreting reliability scores in the context of mental health variables, as inconsistency in mental health self-reporting may reflect clinically meaningful variation, such as stigma-related underreporting, rather than pure measurement noise (Bharadwaj et al., 2015).

Several directions remain open for future work. For reliability scores, incorporating directionality, for instance, by separately tracking underestimation and overestimation patterns, would enable richer characterization of systematic reporting biases. Exploring variable-level normalization during raw score construction is an important avenue to prevent conditions with naturally high onset age variance from disproportionately dominating the participant-level reliability metric. Although our preliminary sensitivity analysis in Appendix O suggests participant rankings are highly robust to this choice for our dataset, formalizing variance-equalized scoring could benefit cohorts with more extreme disease-specific reporting variations. To mitigate the data loss due to stratification, one should investigate utilizing the continuous reliability score directly as a regression weight or covariate; this is non-trivial because the current score is a quantile-normalized rank rather than a calibrated error magnitude. Broadening the reliability scoring framework to capture other forms of longitudinal inconsistency, such as disease status reversals across waves (see Appendix P), is another important direction. The Bayesian adjustment framework could also be extended to handle more than two survey waves, enabling more flexible and accurate adjustment as longitudinal studies accumulate additional follow-up data. While the ablation study for $re_{21}$ indicates a clear benefit to adjusting multiple variables, confirming whether this monotonic improvement holds consistently across other downstream tasks with higher-dimensional features remains for future work. Finally, both methods are currently validated on a single Canadian general-population cohort, and performance may vary across studies with different survey gap lengths, numbers of conditions, or population characteristics. Validation on additional longitudinal cohorts, such as the UK Biobank (Biobank, 2014), is an important next step toward establishing the generalizability of these approaches.

## References

Alessandra T Andreacchi, Alberto Brini, Edwin Van den Heuvel, Graciela Muniz-Terrera, Alexandra Mayhew, Philip St John, Lucy E Stirland, and Lauren E Griffith. An exploration of methods to resolve inconsistent self-reporting of chronic conditions and impact on multimorbidity in the canadian longitudinal study on aging. *Journal of Aging and Health*, 37(1-2):40–53, 2025.

William G Axinn and Stephanie Chardoul. Improving reports of health risks: Life history calendars and measurement of potentially traumatic experiences. *International Journal of Methods in Psychiatric Research*, 30(1):e1853, 2021.

Cátia Azevedo, Sérgio Vilarinho, Ana Sousa Menezes, Fernando Milhazes Mar, and Luís Dias. Vestibular and cochlear dysfunction in aging: Two sides of the same coin? *World Journal of Otorhinolaryngology-Head and Neck Surgery*, 8(04):308–314, 2022.

Melissa J Azur, Elizabeth A Stuart, Constantine Frangakis, and Philip J Leaf. Multiple imputation by chained equations: what is it and how does it work? *International journal of methods in psychiatric research*, 20(1):40–49, 2011.

Gustavo Batista. A study of k-nearest neighbour as an imputation method. *Soft Computing Systems: Design,...*, 2002.

Katy Benjamin, Margaret K Vernon, Donald L Patrick, Eleanor Perfetto, Sandra Nestler-Parr, and Laurie Burke. Patient-reported outcome and observer-reported outcome assessment in rare disease clinical trials: an ispor coa emerging good practices task force report. *Value in Health*, 20(7):838–855, 2017.

Prashant Bharadwaj, Mallesh M Pai, and Agne Suziedelyte. Mental health stigma. Technical report, National Bureau of Economic Research, 2015.

UK Biobank. About uk biobank, 2014.

Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008, 2008.

Benjamin M Bolstad, Rafael A Irizarry, Magnus Åstrand, and Terence P. Speed. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19(2):185–193, 2003.

Richard H Byrd, Peihuang Lu, Jorge Nocedal, and Ciyou Zhu. A limited memory algorithm for bound constrained optimization. *SIAM Journal on scientific computing*, 16(5):1190–1208, 1995.

Raymond J Carroll and Leonard A Stefanski. Approximate quasi-likelihood estimation in models with surrogate predictors. *Journal of the American Statistical Association*, 85 (411):652–663, 1990.

Edward Joseph Caruana, Marius Roman, Jules Hernández-Sánchez, and Piergiorgio Solli. Longitudinal studies. *Journal of thoracic disease*, 7(11):E537, 2015.

Avshalom Caspi, Terrie E Moffitt, Arland Thornton, Deborah Freedman, James W Amell, Honalee Harrington, Judith Smeijers, and Phil A Silva. The life history calendar: a research and clinical assessment method for collecting retrospective event-history data. *International journal of methods in psychiatric research*, 1996.

Jacob Cohen. *Statistical power analysis for the behavioral sciences*. routledge, 2013.

Harald Cramér. *Mathematical methods of statistics*, volume 9. Princeton university press, 1999.

Michael J Domanski, Xin Tian, Colin O Wu, Jared P Reis, Amit K Dey, Yuan Gu, Lihui Zhao, Sejong Bae, Kiang Liu, Ahmed A Hasan, et al. Time course of ldl cholesterol exposure and cardiovascular disease event risk. *Journal of the American College of Cardiology*, 76(13):1507–1516, 2020.

Lorah D Dorn, Lisa M Sontag-Padilla, Stephanie Pabst, Abbigail Tissot, and Elizabeth J Susman. Longitudinal reliability of self-reported age at menarche in adolescent girls: variability across time and setting. *Developmental Psychology*, 49(6):1187, 2013.

Lindsay A Farrer, Louis P Florio, Martha L Bruce, Philip J Leaf, and Myrna M Weissman. Reliability of self-reported age at onset of major depression. *Journal of psychiatric research*, 23(1):35–47, 1989.

Vasudha S Garg, Mihir H Sojitra, Tyagi J Ubhadiya, Nidhi Dubey, Karan Shah, Siddharth Kamal Gandhi, and Priyansh Patel. Understanding the link between adult asthma and coronary artery disease: a narrative review. *Cureus*, 15(8), 2023.

Andrew Gelman, John B Carlin, Hal S Stern, and Donald B Rubin. *Bayesian data analysis*. Chapman and Hall/CRC, 1995.

Mark D Grant, Zdzisiaw H Piotrowski, and Rick Chappell. Self-reported health and survival in the longitudinal study of aging, 1984–1986. *Journal of clinical epidemiology*, 48(3): 375–387, 1995.

Scott M Grundy, James I Cleeman, Stephen R Daniels, Karen A Donato, Robert H Eckel, Barry A Franklin, David J Gordon, Ronald M Krauss, Peter J Savage, Sidney C Smith Jr, et al. Diagnosis and management of the metabolic syndrome: an american heart association/national heart, lung, and blood institute scientific statement. *Circulation*, 112(17): 2735–2752, 2005.

John A Hartigan and Manchek A Wong. Algorithm as 136: A k-means clustering algorithm. *Journal of the royal statistical society. series c (applied statistics)*, 28(1):100–108, 1979.

Andrew C Harvey. Estimating regression models with multiplicative heteroscedasticity. *Econometrica: journal of the Econometric Society*, pages 461–465, 1976.

Friedrich Ihler, Tina Brzoska, Reyhan Altindal, Oliver Dziemba, Henry Völzke, Chia-Jung Busch, and Till Ittermann. Prevalence and risk factors of self-reported hearing loss, tinnitus, and dizziness in a population-based sample from rural northeastern germany. *Scientific Reports*, 14(1):17739, 2024.

Nicholas C Jacobson and Michelle G Newman. Anxiety and depression as bidirectional risk factors for one another: A meta-analysis of longitudinal studies. *Psychological bulletin*, 143(11):1155, 2017.

Timothy P Johnson and Joshua A Mott. The reliability of self-reported age of onset of tobacco, alcohol and illicit drug use. *Addiction*, 96(8):1187–1198, 2001.

Adeel Khoja, Prabha H Andraweera, Zohra S Lassi, Anna Ali, Mingyue Zheng, Maleesa M Pathirana, Emily Aldridge, Melanie R Wittwer, Debajyoti D Chaudhuri, Rosanna Tavella, et al. Risk factors for early-onset versus late-onset coronary heart disease (chd): systematic review and meta-analysis. *Heart, Lung and Circulation*, 32(11):1277–1311, 2023.

Dieter Kraft. A software package for sequential quadratic programming. *Forschungsbericht-Deutsche Forschungs- und Versuchsanstalt fur Luft- und Raumfahrt*, 1988.

Roderick JA Little and Donald B Rubin. *Statistical analysis with missing data*. John Wiley & Sons, 2019.

Elan D Louis, Diep Nguyen, and Ali Ghanem. Reliability of self-report data on age of onset in essential tremor: Data from a prospective, longitudinal cohort. *Neuroepidemiology*, 57 (1):7–13, 2023.

George L Maddox and Elizabeth B Douglass. Self-assessment of health: a longitudinal study of elderly subjects. *Journal of health and social behavior*, pages 87–93, 1973.

Rahul Mazumder, Trevor Hastie, and Robert Tibshirani. Spectral regularization algorithms for learning large incomplete matrices. *The Journal of Machine Learning Research*, 11: 2287–2322, 2010.

Masatoshi Miura, Fumiyuki Goto, Yozo Inagaki, Yasuyuki Nomura, Takeshi Oshima, and Nagisa Sugaya. The effect of comorbidity between tinnitus and dizziness on perceived handicap, psychological distress, and quality of life. *Frontiers in neurology*, 8:722, 2017.

Stephen G Nash. Newton-type minimization via the lanczos method. *SIAM Journal on Numerical Analysis*, 21(4):770–788, 1984.

Gilbert R Parra, Susan E O'Neill, and Kenneth J Sher. Reliability of self-reported age of substance involvement onset. *Psychology of Addictive Behaviors*, 17(3):211, 2003.

Bruce M Psaty, Curt D Furberg, Lewis H Kuller, Mary Cushman, Peter J Savage, David Levine, Daniel H O'Leary, R Nick Bryan, Melissa Anderson, and Thomas Lumley. Association between blood pressure level and the risk of myocardial infarction, stroke, and total mortality: the cardiovascular health study. *Archives of internal medicine*, 161(9): 1183–1192, 2001.

Ariel E Quinio and TC Lam. Methods and biases in measuring change with self-reports. *Basic elements of survey research in education: Addressing the problems your advisor never told you about*, pages 193–217, 2021.

Karen G Raphael and Joseph J Marbach. When did your pain start?: reliability of self-reported age of onset of facial pain. *The Clinical journal of pain*, 13(4):352–359, 1997.

Lisa Reisinger, Fabian Schmidt, Kaja Benz, Lorenzo Vignali, Sebastian Roesch, Martin Kronbichler, and Nathan Weisz. Ageing as risk factor for tinnitus and its complex interplay with hearing loss—evidence from online and nhanes data. *BMC medicine*, 21(1): 283, 2023.

Stephen Rhodes, Nathaniel R Greene, and Moshe Naveh-Benjamin. Age-related differences in recall and recognition: A meta-analysis. *Psychonomic Bulletin & Review*, 26(5):1529–1547, 2019.

Maria Luísa Rocha, Dinis Pestana, and António Gomes de Menezes. Heavy tails and mixtures of normal random variables. *CEEAplA-A-Working Paper Series*, pages 1–10, 2012.

Audrey M Shillington, Mark B Reed, and John D Clapp. Self-report stability of adolescent cigarette use across ten years of panel study data. *Journal of Child & Adolescent Substance Abuse*, 19(2):171–191, 2010.

D Skybova, H Slachtova, H Tomaskova, J Klanova, and R Madar. The age of chronic diseases onset–the results of the longitudinal study. *European Journal of Public Health*, 32, 2022.

Matthew C Tattersall, Jodi H Barnet, Claudia E Korcarz, Erika W Hagen, Paul E Peppard, and James H Stein. Late-onset asthma predicts cardiovascular disease events: the wisconsin sleep cohort. *Journal of the American Heart Association*, 5(9):e003448, 2016.

Marijke Veenstra and Astri Syse. Health behaviour changes and onset of chronic health problems in later life. *Norsk epidemiologi*, 22(2), 2012.

Chi Wang, Yu Yuan, Mengyi Zheng, AN Pan, Miao Wang, Maoxiang Zhao, Yao Li, Siyu Yao, Shuohua Chen, Shouling Wu, et al. Association of age of onset of hypertension with cardiovascular diseases and mortality. *Journal of the American College of Cardiology*, 75 (23):2921–2930, 2020.

Susanne Wiegand, Reinhard Berner, Antonius Schneider, Ellen Lundershausen, and Andreas Dietz. Otitis externa: investigation and evidence-based treatment. *Deutsches Ärzteblatt International*, 116(13):224, 2019.

Robert F Woolson. Wilcoxon signed-rank test. *Wiley encyclopedia of clinical trials*, pages 1–3, 2007.

## Appendix A. Descriptions of 55 onset age variables collected at enrollment and follow-up in CanPath data

Table 4: Full names and condition descriptions for the onset age pairs at enrollment and at follow-up.

| **Enrolment Variable** | **Follow-up Variable** | **Onset age for** |
|---|---|---|
| WH_CONTRACEPTIVES_AGE | F1_WH_CONTRACEPTIVES_AGE | Hormonal contraceptives use |
| WH_MENOPAUSE_AGE | F1_WH_MENOPAUSE_AGE | Menopause |
| WH_HRT_AGE | F1_WH_HRT_AGE | Hormone replacement therapy |
| WH_HYSTERECTOMY_AGE | F1_WH_HYSTERECTOMY_AGE | Hysterectomy |
| DIS_HBP_AGE | F1_DIS_HBP_AGE | High blood pressure or hypertension |
| DIS_MI_AGE | F1_DIS_MI_AGE | Myocardial infarction |
| DIS_ASTHMA_AGE | F1_DIS_ASTHMA_AGE | Asthma |
| DIS_EMPHYSEMA_AGE | F1_DIS_EMPHYSEMA_AGE | Emphysema |
| DIS_CB_AGE | F1_DIS_CB_AGE | Chronic bronchitis |
| DIS_COPD_AGE | F1_DIS_COPD_AGE | COPD |
| DIS_DEP_AGE | F1_DIS_DEP_AGE | Major depression |
| DIS_LC_AGE | F1_DIS_LC_AGE | Liver cirrhosis |
| DIS_CH_AGE | F1_DIS_CH_AGE | Chronic hepatitis |
| DIS_CROHN_AGE | F1_DIS_CROHN_AGE | Crohn's disease |
| DIS_UC_AGE | F1_DIS_UC_AGE | Ulcerative colitis |
| DIS_IBS_AGE | F1_DIS_IBS_AGE | Irritable bowel syndrome |
| DIS_ECZEMA_AGE | F1_DIS_ECZEMA_AGE | Eczema |
| DIS_SLE_AGE | F1_DIS_SLE_AGE | Systemic lupus erythematosus |
| DIS_PS_AGE | F1_DIS_PS_AGE | Psoriasis |
| DIS_MS_AGE | F1_DIS_MS_AGE | Multiple sclerosis |
| DIS_OP_AGE | F1_DIS_OP_AGE | Osteoporosis |
| DIS_ARTHRITIS_AGE | F1_DIS_ARTHRITIS_AGE | Arthritis |
| DIS_CANCER1_AGE | F1_DIS_CANCER1_AGE | Breast cancer |
| DIS_CANCER2_AGE | F1_DIS_CANCER2_AGE | Colon cancer |
| DIS_CANCER3_AGE | F1_DIS_CANCER3_AGE | Bronchus and lung cancer |
| SMK_CIG_DAILY_CUR_ONSET | F1_SMK_CIG_DAILY_CUR_ONSET | Daily cigarette smoking |
| DIS_ENDO_TD_AGE | F1_DIS_TD_AGE | Thyroid disease |
| DIS_ENDO_HB_CHOL_AGE | F1_DIS_HIGH_CHOL_AGE | High blood cholesterol |
| DIS_CARDIO_ATRIAL_AGE | F1_DIS_ATRIAL_AGE | Atrial fibrillation |
| DIS_CARDIO_ANGINA_AGE | F1_DIS_ANGINA_AGE | Angina |
| DIS_CARDIO_VHD_AGE | F1_DIS_VHD_AGE | Valvular heart disease |
| DIS_RESP_SLEEP_APNEA_AGE | F1_DIS_SLEEP_APNEA_AGE | Sleep apnea |

| **Enrolment Variable** | **Follow-up Variable** | **Description** |
|---|---|---|
| DIS_GASTRO_ULCERS_AGE | F1_DIS_ULCERS_AGE | Stomach or duodenal ulcers |
| DIS_GASTRO_GERD_AGE | F1_DIS_GERD_AGE | Persistent acid reflux (GERD) |
| DIS_LIVER_FATTY_AGE | F1_DIS_FATTY_AGE | Fatty liver (NAFLD/NASH) |
| DIS_LIVER_PANCREATITIS_AGE | F1_DIS_PANCREAT_AGE | Pancreatitis |
| DIS_MH_BIPOLAR_AGE | F1_DIS_BIPOLAR_DIS_AGE | Bipolar disorder |
| DIS_MH_PTSD_AGE | F1_DIS_PTSD_AGE | Post-traumatic stress disorder |
| DIS_MH_SCHIZOPHRENIA_AGE | F1_DIS_SCHIZOPHRENIA_AGE | Schizophrenia |
| DIS_MH_OCD_AGE | F1_DIS_OCD_AGE | Obsessive compulsive disorder |
| DIS_MH_ANXIETY_AGE | F1_DIS_ANXIETY_DIS_AGE | Anxiety disorder |
| DIS_MH_EATING_AGE | F1_DIS_EATING_DIS_AGE | Eating disorder |
| DIS_MH_ADDICTION_AGE | F1_DIS_ADDICTION_DIS_AGE | Addiction disorder |
| DIS_NEURO_MIGRAINE_AGE | F1_DIS_MIGRAINE_AGE | Migraines |
| DIS_NEURO_EPILEPSY_AGE | F1_DIS_EPILEPSY_AGE | Epilepsy or seizures |
| DIS_NEURO_PARKINSON_AGE | F1_DIS_PARKINSON_AGE | Parkinson's disease |
| DIS_BONE_FIBROMYALGIA_AGE | F1_DIS_FIBROMYAL_AGE | Fibromyalgia |
| DIS_INFEC_HIV_AGE | F1_DIS_HIV_AGE | HIV |
| DIS_INFEC_HPV_AGE | F1_DIS_HPV_AGE | Genital warts (HPV infection) |
| DIS_INFEC_HERPES_AGE | F1_DIS_HERPES_AGE | Genital herpes |
| DIS_EYE_MACULAR_AGE | F1_DIS_MACULAR_AGE | Macular degeneration |
| DIS_EYE_GLAUCOMA_AGE | F1_DIS_GLAUCOMA_AGE | Glaucoma |
| DIS_EYE_CATARACTS_AGE | F1_DIS_CATARACT_AGE | Cataracts |
| DIS_EAR_TINNITUS_AGE | F1_DIS_TINNITUS_AGE | Tinnitus |
| DIS_EAR_LOSS_AGE | F1_DIS_HEARING_LOSS_AGE | Hearing loss |

## Appendix B. Derivation of the posterior distribution

We assume a diffuse (non-informative) Normal prior on the latent true value:

$$X_{ij}^* \sim \mathcal{N}(\mu_0, \sigma_0^2),$$

where $\mu_0$ represents our prior mean belief about the true onset age and $\sigma_0^2$ represents our prior uncertainty. In the absence of strong prior knowledge about the true onset age, we use a diffuse prior by letting $\sigma_0^2 \to \infty$. By Bayes' theorem and the conditional independence of $X_{ij}^{(e)}$ and $X_{ij}^{(f)}$ given $X_{ij}^*$, the posterior is:

$$p(X_{ij}^* \mid X_{ij}^{(e)}, X_{ij}^{(f)}) \propto p(X_{ij}^{(e)} \mid X_{ij}^*) \cdot p(X_{ij}^{(f)} \mid X_{ij}^*) \cdot p(X_{ij}^*).$$

The posterior distribution is Normal due to conjugacy: when both the likelihood and prior are Normal distributions, the posterior is also Normal. Taking the logarithm:

$$\begin{aligned}\log p(X_{ij}^* \mid X_{ij}^{(e)}, X_{ij}^{(f)}) &\propto -\frac{(X_{ij}^{(e)} - X_{ij}^*)^2}{2{\sigma_j^{(e)}}^2} - \frac{(X_{ij}^{(f)} - X_{ij}^*)^2}{2{\sigma_j^{(f)}}^2} - \frac{(X_{ij}^* - \mu_0)^2}{2\sigma_0^2} \\ &= -\frac{1}{2}\left[\frac{(X_{ij}^{(e)} - X_{ij}^*)^2}{{\sigma_j^{(e)}}^2} + \frac{(X_{ij}^{(f)} - X_{ij}^*)^2}{{\sigma_j^{(f)}}^2} + \frac{(X_{ij}^* - \mu_0)^2}{\sigma_0^2}\right].\end{aligned}$$

As $\sigma_0^2 \to \infty$, the prior term vanishes, leaving:

$$\log p(X_{ij}^* \mid X_{ij}^{(e)}, X_{ij}^{(f)}) \propto -\frac{1}{2}\left[\frac{(X_{ij}^{(e)} - X_{ij}^*)^2}{{\sigma_j^{(e)}}^2} + \frac{(X_{ij}^{(f)} - X_{ij}^*)^2}{{\sigma_j^{(f)}}^2}\right].$$

Expanding the squared terms:

$$\begin{aligned}&-\frac{1}{2}\left[\frac{{X_{ij}^{(e)}}^2 - 2X_{ij}^{(e)}X_{ij}^* + {X_{ij}^*}^2}{{\sigma_j^{(e)}}^2} + \frac{{X_{ij}^{(f)}}^2 - 2X_{ij}^{(f)}X_{ij}^* + {X_{ij}^*}^2}{{\sigma_j^{(f)}}^2}\right] \\ &= -\frac{1}{2}\left[\left(\frac{1}{{\sigma_j^{(e)}}^2} + \frac{1}{{\sigma_j^{(f)}}^2}\right){X_{ij}^*}^2 - 2\left(\frac{X_{ij}^{(e)}}{{\sigma_j^{(e)}}^2} + \frac{X_{ij}^{(f)}}{{\sigma_j^{(f)}}^2}\right)X_{ij}^* + \text{const}\right].\end{aligned}$$

Completing the square, we obtain a Normal distribution with:

$$\text{Posterior variance:} \quad \tau_{ij}^2 = \left(\frac{1}{{\sigma_j^{(e)}}^2} + \frac{1}{{\sigma_j^{(f)}}^2}\right)^{-1},$$

$$\text{Posterior mean:} \quad \mu_{ij} = \tau_{ij}^2\left(\frac{X_{ij}^{(e)}}{{\sigma_j^{(e)}}^2} + \frac{X_{ij}^{(f)}}{{\sigma_j^{(f)}}^2}\right).$$

Therefore, the posterior distribution is:

$$X_{ij}^* \mid X_{ij}^{(e)}, X_{ij}^{(f)}, a_i^{(e)}, \Delta_i \sim \mathcal{N}(\mu_{ij}, \tau_{ij}^2).$$

## Appendix C. Synthetic experiment details for reliability score-based stratification

We generated synthetic data following a participant-specific noise model that mimics the structure of real healthcare data: Let $\mathbf{X}^{(e)} \in \mathbb{R}^{n\times p}$ and $\mathbf{X}^{(f)} \in \mathbb{R}^{n\times p}$ denote enrollment and follow-up responses. For participant $i$:

$$\begin{aligned} X_{ij}^{(e)} &= (D_{\text{true}})_{ij} + \varepsilon_{ij}^{(e)}, \quad \text{where } \varepsilon_{ij}^{(e)} \sim \mathcal{N}(0, \lambda_i^2) \\ X_{ij}^{(f)} &= (D_{\text{true}})_{ij} + \varepsilon_{ij}^{(f)}, \quad \text{where } \varepsilon_{ij}^{(f)} \sim \mathcal{N}(0, \lambda_i^2) \\ D_{ij} &= X_{ij}^{(f)} - X_{ij}^{(e)} = 2(D_{\text{true}})_{ij} + \delta_{ij}, \quad \text{where } \delta_{ij} \sim \mathcal{N}(0, 2\lambda_i^2) \end{aligned}$$

Each participant $i$ has a noise parameter $\lambda_i \sim \text{Uniform}(1.5, 2.5)$, serving as ground-truth data quality. Signal is scaled (std = 8) to ensure strong signal-to-noise ratio. Missingness is correlated with noise level (high-$\lambda$ participants have higher missing probability). We tested 60 configurations:

- Sample sizes: $n \in \{100, 200, 300, 400\}$
- Variables: $p \in \{20, 30, 40\}$
- Missing rates: $0.5, 0.6, 0.7, 0.8, 0.9$
- Each configuration evaluated with 3 independent random seeds

The evaluation metrics include:

- Rank alignment (primary metric): The Spearman correlation $|\rho|$ between reliability scores and ground-truth noise $\lambda$. Higher correlation indicates scores better rank participants by data quality. Expected sign: negative (higher reliability $\rightarrow$ lower noise), but we report absolute value for interpretability.
- Discrimination: Kruskal-Wallis $H$-statistic ($-\log_{10} p$) testing whether reliability score quartile tiers differ in $\lambda$. Higher values indicate cleaner separation by data quality.
- Consistency: Stability of tier assignment under subsampling: fraction of participants assigned the same quartile across 5 bootstrap resamples (80% of data). Higher = more stable stratification.

The pipeline variants include

1. Full method (ours): the complete pipeline of §3.1 (SoftImpute + simple averaging).
2. No imputation: SoftImpute removed. Averaging with observed values only.
3. Mean imputation: SoftImpute replaced with mean imputation.
4. kNN imputation: SoftImpute replaced with kNN imputation.
5. MICE: SoftImpute replaced with Multivariate Imputation by Chained Equations.

6. Variance-based weighting: simple averaging replaced with variance-based weighting.

The results are as follows.

Table 5: Rank alignment

| Method/Missingness | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 | Average |
|---|---|---|---|---|---|---|
| (1) SoftImpute+Avg | 0.808 | 0.831 | 0.839 | 0.827 | 0.841 | 0.829 |
| (2) No Imputation | 0.372 | 0.334 | 0.262 | 0.266 | 0.216 | 0.290 |
| (3) Mean Imputation | 0.808 | 0.830 | 0.838 | 0.826 | 0.840 | 0.828 |
| (4) kNN Imputation | 0.346 | 0.431 | 0.478 | 0.494 | 0.527 | 0.455 |
| (5) MICE Imputation | 0.808 | 0.830 | 0.838 | 0.826 | 0.840 | 0.828 |
| (6) Variance-Based | 0.798 | 0.822 | 0.830 | 0.815 | 0.828 | 0.819 |

Table 6: Discrimination

| Method/Missingness | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 | Average |
|---|---|---|---|---|---|---|
| (1) SoftImpute+avg | 29.318 | 25.545 | 21.811 | 17.981 | 16.316 | 22.194 |
| (2) No imputation | 8.706 | 6.784 | 5.276 | 4.133 | 3.353 | 5.650 |
| (3) Mean imputation | 29.241 | 25.511 | 21.783 | 17.899 | 16.341 | 22.155 |
| (4) kNN imputation | 5.958 | 7.055 | 6.629 | 6.023 | 6.111 | 6.355 |
| (5) MICE imputation | 29.241 | 25.511 | 21.783 | 17.899 | 16.341 | 22.155 |
| (6) Variance-based | 28.823 | 25.047 | 21.487 | 17.585 | 15.963 | 21.781 |

Table 7: Consistency

| Method/Missingness | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 | Average |
|---|---|---|---|---|---|---|
| (1) SoftImpute+avg | 0.961 | 0.957 | 0.952 | 0.944 | 0.945 | 0.952 |
| (2) No imputation | 0.961 | 0.956 | 0.953 | 0.946 | 0.941 | 0.951 |
| (3) Mean imputation | 0.960 | 0.956 | 0.951 | 0.944 | 0.946 | 0.951 |
| (4) kNN imputation | 0.959 | 0.957 | 0.953 | 0.947 | 0.943 | 0.952 |
| (5) MICE imputation | 0.960 | 0.956 | 0.951 | 0.944 | 0.946 | 0.951 |
| (6) Variance-based | 0.961 | 0.956 | 0.951 | 0.943 | 0.945 | 0.951 |

Our comprehensive ablation study demonstrates that SoftImpute+avg is the best strategy for reliability scoring in the presence of missing data across the three metrics. On the most important metric, rank alignment, SoftImpute+avg achieves an average score of 0.829, outperforming mean imputation (0.828, +0.12%) and MICE imputation (0.828, +0.12%), while substantially surpassing variance-based weighting (0.819, +1.2%), kNN imputation (0.455, +82%), and no imputation (0.290, +186%). Note that the differences between (1) and (5) are small; we chose SoftImpute primarily for its computational efficiency over the iterative MICE procedure.

## Appendix D. Synthetic experiment details for Bayesian adjustment

We conducted comprehensive synthetic data experiments to validate the Bayesian adjustment method for estimating true latent onset ages from inconsistent measurements. The experiments generated synthetic data under the exact measurement error model assumptions specified in Section 3.2, where enrollment and follow-up measurements are treated as noisy observations of a latent true age value. We have the following data generation scheme.

1. Generated true latent onset ages: $X_{ij}^* \sim \mathcal{N}(35, 10)$ for each participant and variable.
2. Simulated enrollment and follow-up measurements as noisy observations using Equation 2.
3. Fitted the Bayesian model via maximum likelihood estimation to estimate variance parameters $\{\sigma_0^2, \alpha_0, \alpha_1, \delta_0, \delta_1\}$.
4. Applied Bayesian adjustment to compute posterior mean estimates $\hat{X}_{ij}^*$:
5. Evaluated four methods on each variable: enrollment-only, follow-up-only, simple average, and Bayesian adjustment.

The seven scenarios in the following table were carefully designed to test the method under different realistic conditions.

Table 8: Scenario specifications for synthetic experiments.

| **Scenario** | **N** | $\boldsymbol{\sigma_0}$ | $\boldsymbol{\alpha_0}$ | $\boldsymbol{\alpha_1}$ | $\boldsymbol{\delta_0}$ | $\boldsymbol{\delta_1}$ | **Age range** | **Gap range** |
|---|---|---|---|---|---|---|---|---|
| 1. Baseline[a] | 2,000 | 1.0 | −3.0 | 0.010 | 0.10 | 0.010 | 20–70 | 2–20 yr |
| 2. Low error[b] | 2,000 | 0.5 | −4.0 | 0.005 | 0.05 | 0.005 | 20–70 | 2–20 yr |
| 3. High error[c] | 2,000 | 2.0 | −2.0 | 0.020 | 0.20 | 0.020 | 20–70 | 2–20 yr |
| 4. Age-dependent[d] | 2,000 | 1.0 | −3.0 | 0.050 | 0.10 | 0.010 | 20–70 | 2–20 yr |
| 5. Gap-dependent[e] | 2,000 | 1.0 | −3.0 | 0.010 | 0.10 | 0.050 | 20–70 | 2–20 yr |
| 6. Long follow-up[f] | 2,000 | 1.0 | −3.0 | 0.010 | 0.10 | 0.010 | 20–70 | 10–30 yr |
| 7. Large sample[g] | 10,000 | 1.0 | −3.0 | 0.010 | 0.10 | 0.010 | 20–70 | 2–20 yr |
| 8. Directional bias[h] | 2,000 | 1.0 | -3.0 | 0.010 | 0.10 | 0.010 | 20–70 | 2–20 yr |
| 9. Digit heaping[i] | 2,000 | 1.0 | -3.0 | 0.010 | 0.10 | 0.010 | 20–70 | 2–20 yr |
| 10. Outlier contamination[j] | 2,000 | 1.0 | -3.0 | 0.010 | 0.10 | 0.010 | 20–70 | 2–20 yr |

[a] Standard configuration with moderate measurement error.
[b] Minimal noise; tests method when baseline error is small.
[c] Substantial noise; tests method when baseline error is large.
[d] Strong age effect (5× baseline); enrollment error increases steeply with age.
[e] Strong gap effect (5× baseline); follow-up error increases steeply over time.
[f] Extended follow-up intervals; accumulating measurement error over time.
[g] 5× larger sample; tests stability and precision with increased sample size.
[h] Recall bias: enrollment under-reports by 1 yr and follow-up over-reports by 2 yr on average; tests whether variance-weighted posterior can recover truth when both observations are biased in opposite directions.
[i] 30% of reports rounded to the nearest multiple of 5, mimicking the "around 40-ish" recall heuristic documented in the self-report literature.
[j] 5% of reports suffer a digit transposition (e.g., 56 → 65), producing excess kurtosis; tests whether per-variable variance estimation and posterior weighting remain calibrated under heavy-tailed contamination.

We benchmarked our Bayesian adjustment method using enrollment data only, follow-up data only and the average of enrollment and follow-up data. The results are in the following table.

Table 9: MAE for Bayesian adjustment and three benchmark methods across seven synthetic scenarios. The improvement range shows how much Bayesian adjustment reduces MAE compared to the best-performing benchmark and the worst-performing benchmark.

| **Scenario** | **Enrollment** | **Follow-up** | **Average** | **Bayesian** | **Improvement** |
|---|---|---|---|---|---|
| 1. Baseline | 0.2239 | 0.2495 | 0.1674 | **0.1674** | 0.00% to +32.94% |
| 2. Low error | 0.0603 | 0.0634 | 0.0438 | **0.0437** | +0.23% to +31.07% |
| 3. High error | 0.9191 | 1.1493 | 0.7352 | **0.7225** | +1.73% to +37.15% |
| 4. Age-dependent | 0.5886 | 0.6582 | 0.4432 | **0.4426** | +0.14% to +32.78% |
| 5. Gap-dependent | 0.2244 | 0.3116 | 0.1922 | **0.1822** | +5.20% to +41.52% |
| 6. Long follow-up | 0.2252 | 0.2586 | 0.1712 | **0.1708** | +0.23% to +34.01% |
| 7. Large sample | 0.2241 | 0.2485 | 0.1667 | **0.1665** | +0.12% to +32.97% |
| 8. Directional bias | 1.0023 | 1.9990 | 0.4995 | **0.4504** | +9.83% to +77.47% |
| 9. Digit heaping | 0.5384 | 0.5526 | 0.4525 | **0.4539** | −0.31% to +17.86% |
| 10. Outlier contamination | 1.4535 | 1.5743 | 0.4386 | **1.4243** | +0.99% to +9.53% |
| **Average** | **0.5460** | **0.7065** | **0.4310** | **0.4224** | **+2.00% to +40.21%** |

## Appendix E. SDoH comparison between included and excluded participants

We conducted a demographic comparison between included ($n$ = 25,018) and excluded ($n$ = 72,390) participants. We reported effect sizes as the primary criterion The results are summarized in Table 10.

Table 10: Demographic comparison between included and excluded participants.

| **Variable** | **Effect size** |
|---|---|
| Age | $d = 0.53$ (medium) |
| Sex | $V = 0.34$ (medium) |
| Education level | $V = 0.10$ (negligible) |
| Household income | $V = 0.10$ (negligible) |

Two of four variables show negligible or small effect sizes. The two medium effects both have structural explanations rooted in questionnaire design:

- Age ($d = 0.53$): Older participants have had more time to develop multiple conditions, so they accumulate more non-missing onset age entries and are more likely to clear the $\rho = 53/55$ threshold. Since $d$ is positive, we include more older people; this age skew does not undermine the method's utility because reliability score-based stratification is intended for risk assessment in populations with multi-condition histories, so applicability to older participants is expected.
- Sex ($V = 0.34$): Four of the 55 variables are female-specific (hormonal contraceptives, menopause, HRT, hysterectomy), so female participants are eligible to report onset ages for more variables and are more likely to pass the threshold.

## Appendix F. The distribution of reliability scores before quantile normalization

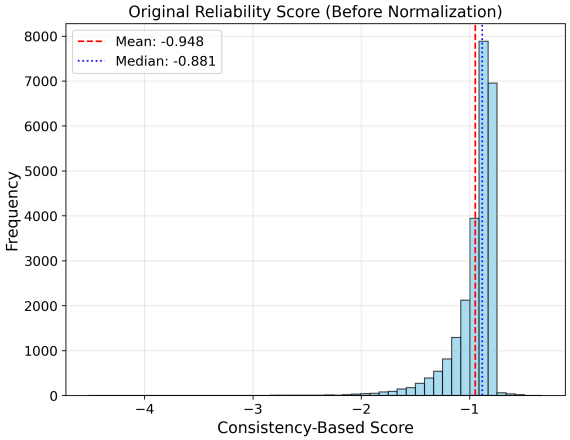


Figure 4: The original reliability scores before quantile normalization.

## Appendix G. Sensitivity analysis of Louvain resolution parameter for disease clustering

To verify that the superior biological coherence observed in the high-reliability cohort's correlation network is robust and not an artifact of the specific Louvain resolution parameter, we conducted a sensitivity analysis. We swept the resolution parameter across $\gamma \in \{1.0, 1.5, 2.0\}$. In all settings, the natural correlation threshold was kept at 0.0 (i.e., all positive correlation edges were retained in the network to avoid arbitrary sparsification). We evaluated the resulting network partition quality using our two primary biological coherence metrics: the dominant-category percentage and the cluster entropy. The empirical results are summarized in Table 11 and Table 12.

Table 11: Sensitivity analysis of the Dominant Category Percentage across varying Louvain resolution parameters for high- and low-reliability cohorts (correlation threshold = 0.0).

| **Threshold** | **Resolution** | **Dominant Category %** | | |
|---|---|---|---|---|
| | | **High** | **Low** | **Difference** |
| 0.0 | 1.0 | 38.5% | 32.0% | +6.54% |
| 0.0 | 1.5 | 57.3% | 49.9% | +7.39% |
| 0.0 | 2.0 | 60.9% | 54.1% | +6.84% |
| **mean** | | **52.2%** | **45.4%** | **+6.92%** |

Table 12: Sensitivity analysis of Cluster Entropy across varying Louvain resolution parameters for high- and low-reliability cohorts (correlation threshold = 0.0).

| **Threshold** | **Resolution** | **Cluster Entropy** | | |
|---|---|---|---|---|
| | | **High** | **Low** | **Difference** |
| 0.0 | 1.0 | 2.21 | 2.32 | −0.11 |
| 0.0 | 1.5 | 1.19 | 1.40 | −0.21 |
| 0.0 | 2.0 | 0.94 | 1.15 | −0.21 |
| **mean** | | **1.45** | **1.62** | −0.17 |

Note that the entropy is computed within each cluster, not across clusters. Lower entropy indicates greater coherence with the taxonomy defined in Appendix H, not necessarily better graph-structural separation. We make coherence the primary outcome because the scientific question is whether the recovered clusters are biologically meaningful, an external-validation question against an independent ground-truth labeling. Internal graph-structural metrics, by contrast, only ask whether the partition is consistent with the graph it was computed from, and in our setting, they can be inflated by spurious high-correlation edges without biological basis. Across every evaluated configuration, the high-reliability cohort consistently groups diseases into biologically cohesive categories more effectively than the low-reliability cohort, demonstrating an average improvement of +6.92% in dominant-category representation. In addition, the high-reliability cohort consistently yields lower entropy across its clusters, with an average reduction of −0.17, proving that diseases are more systematically grouped with biologically related conditions.

## Appendix H. Classification of diseases

Table 13: Disease categorization rules for 55 onset age variables. Patterns are matched case-insensitively against variable names.

| **Category** | **Number** | **Matching patterns** |
|---|---|---|
| Cardiovascular | 5 | `cardio, mi_, hbp_, angina, atrial` |
| Respiratory | 5 | `asthma, copd, emphysema, resp_, cb_` |
| Gastrointestinal/Hepatic | 9 | `gastro, ibs, crohn, uc_, liver, pancreatitis, ch_, lc_` |
| Mental health | 8 | `mh_, dep_` |
| Musculoskeletal | 3 | `arthritis, bone_, op_` |
| Neurological | 4 | `ms_, neuro_, migraine` |
| Endocrine/Metabolic | 2 | `endo_, fatty_` |
| Sensory (eye/ear) | 5 | `eye_, ear_` |
| Dermatological | 2 | `eczema, ps_` |
| Infectious | 3 | `infec_` |
| Cancer | 3 | `cancer` |
| Autoimmune | 1 | `sle_` |
| Other | 5 | (no match) |
| **Total** | **55** | |

## Appendix I. Coherence metrics

Suppose $C$ is the number of medical categories in the cluster, $n_i$ is the number of diseases in category $i$, and $N$ is the total number of diseases in the cluster. Entropy is

$$H = -\sum_{i=1}^{C} p_i \log_2(p_i), \quad \text{where} \quad p_i = \frac{n_i}{N}.$$

Dominant category percentage is:

$$D = \frac{\max_i(n_i)}{N} \times 100\%.$$

## Appendix J. Cluster composition by Disease Category

The cluster composition for Figure 3 by disease category according to the categorization in Appendix H are as follows.

Table 14: Low-reliability cohort: cluster composition by disease category. Average dominant % = 30.9.

| Cluster | Size | Dominant category (%) | Category distribution |
|---|---|---|---|
| 1 | 11 | Respiratory/GI/Heptic (27%) | Respiratory 3, GI/Hepatic 3, Musculoskeletal 2, Cardiovascular 2, Neurological 1 |
| 2 | 11 | Mental Health (36%) | Mental Health 4, GI/Hepatic 3, Cancer 1, Cardiovascular 1, Endocrine 1, Autoimmune 1 |
| 3 | 8 | Respiratory/Sensory (25%) | Respiratory 2, Sensory 2, Endocrine 1, GI/Hepatic 1, Cardiovascular 1, Musculoskeletal 1 |
| 4 | 8 | Sensory Eye/Ear (37.5%) | Sensory 3, Cancer 2, Infectious 2, Dermatological 1 |
| 0 | 7 | GI/Hepatic/Mental Health (29%) | GI/Hepatic 2, Mental Health 2, Cardiovascular 1, Neurological 1, Dermatological 1 |

Table 15: High-reliability cohort: cluster composition by disease category. Average dominant % = 43.8.

| Cluster | Size | Dominant category (%) | Category distribution |
|---|---|---|---|
| 2 | 11 | Cardiovascular (45%) | Cardiovascular 5, Endocrine 2, Sensory 2, Respiratory 1, Infectious 1 |
| 3 | 10 | GI/Hepatic (60%) | GI/Hepatic 6, Musculoskeletal 3, Autoimmune 1 |
| 1 | 9 | Mental Health (33%) | Mental Health 3, GI/Hepatic 2, Neurological 2, Cancer 1, Respiratory 1 |
| 4 | 8 | Mental Health (37.5%) | Mental Health 3, Respiratory 2, Dermatological 2, GI/Hepatic 1 |
| 0 | 7 | Sensory Eye/Ear (43%) | Sensory 3, Cancer 2, Infectious 1, Respiratory 1 |

## Appendix K. Tie-breaking rules for determining the best model

When selecting the best model for each task, we apply the following tie-breaking rules:

**Reliability score-based stratification**

- **Classification tasks:** Models are ranked by F1 score on each cohort. If multiple models achieve the same F1 score, we compare the their precisions.
- **Regression tasks:** Models are ranked by RMSE on each cohort. If multiple models achieve the same RMSE, we compare their MAE values.

**Bayesian adjustment**

- **Classification tasks:** Models are ranked by F1 score for using enrollment data only, using follow-up data only, and using Bayesian adjustment independently.
- **Regression tasks:** Models are ranked by RMSE for using enrollment data only, using follow-up data only, and using Bayesian adjustment independently. If multiple models achieve the same RMSE, we compare their MAE values.

## Appendix L. SDoH comparison between reliability-stratified cohorts

To assess whether any observed differences in predictive performance between high- and low-reliability cohorts might reflect systematic demographic differences between the cohorts rather than genuine data quality gains, we compared the distribution of SDoH, including age, sex, income, and education, between high- and low-reliability cohorts for each task ($cl_{11}$–$cl_{13}, re_{11} - re_{13}$). We reported effect sizes as the primary criterion in the following table, Cohen's $d$ for continuous variables (negligible $|d| < 0.2$, small $0.2 \leq |d| \leq 0.5$, medium $0.5 \leq |d| \leq 0.8$, large $|d| \geq 0.8$) and Cramér's $V$ for categorical variables (with negligible $|V| < 0.1$, small $0.1 \leq |V| < 0.3$, medium $0.3 \leq |V| \leq 0.5$, large $|V| \geq 0.5$).

| **Task/SDoH variable** | **age** | **sex** | **income** | **education** |
|---|---|---|---|---|
| dep_age | 0.2750 | 0.1237 | -0.0839 | -0.0612 |
| dep_ever | 0.2760 | 0.1201 | -0.0837 | -0.0680 |
| diab_ever | 0.2750 | 0.1228 | -0.0730 | -0.0624 |
| endo_sugar_age | 0.2926 | 0.1273 | -0.0809 | -0.0717 |
| endo_sugar_ever | 0.2962 | 0.1286 | -0.0866 | -0.0779 |
| hearing_loss_age | 0.2248 | 0.0921 | -0.1457 | -0.0532 |

Table 16: SDOH effect sizes by task.

We concluded that across all tasks, SDoH differences are negligible to small.

## Appendix M. Full predictive results for reliability score-based stratification

Table 17: Cross-validation classification performance across over $cl_{11}, cl_{12}$ and $cl_{13}$ for five models across low- and high-reliability cohorts (divided by median reliability score). Performance metrics include precision and recall. F1 scores are used for determining the best model. High-reliability cohort rows are highlighted in gray. The best models for the high and low reliability cohorts are denoted with $(*)$ and $(\star)$, respectively.

| ID | Target | Model | Cohort | Metrics | | |
|---|---|---|---|---|---|---|
| | | | | Precision | Recall | F1 |
| $cl_{11}$ | Depression | Logistic $(\star, *)$ | Low | 0.745 | 0.731 | 0.731 |
| | | | High | 0.733 | 0.719 | 0.706 |
| | | RF | Low | 0.710 | 0.702 | 0.703 |
| | | | High | 0.670 | 0.669 | 0.662 |
| | | DT | Low | 0.717 | 0.703 | 0.704 |
| | | | High | 0.683 | 0.675 | 0.661 |
| | | GB | Low | 0.730 | 0.721 | 0.721 |
| | | | High | 0.708 | 0.698 | 0.685 |
| | | SVM | Low | 0.737 | 0.723 | 0.723 |
| | | | High | 0.744 | 0.718 | 0.700 |
| $cl_{12}$ | High blood sugar | Logistic $(\star, *)$ | Low | 0.928 | 0.928 | 0.922 |
| | | | High | 0.957 | 0.957 | 0.954 |
| | | RF | Low | 0.916 | 0.918 | 0.912 |
| | | | High | 0.949 | 0.951 | 0.948 |
| | | DT | Low | 0.865 | 0.862 | 0.863 |
| | | | High | 0.914 | 0.910 | 0.911 |
| | | GB | Low | 0.921 | 0.922 | 0.916 |
| | | | High | 0.950 | 0.951 | 0.948 |
| | | SVM $(\star, *)$ | Low | 0.928 | 0.928 | 0.922 |
| | | | High | 0.957 | 0.957 | 0.954 |
| $cl_{13}$ | Diabetes | Logistic | Low | 0.589 | 0.767 | 0.666 |
| | | | High | 0.564 | 0.751 | 0.644 |
| | | RF $(\star)$ | Low | 0.681 | 0.719 | 0.693 |
| | | | High | 0.679 | 0.727 | 0.697 |
| | | DT | Low | 0.663 | 0.626 | 0.641 |
| | | | High | 0.700 | 0.671 | 0.674 |
| | | GB $(*)$ | Low | 0.679 | 0.719 | 0.690 |
| | | | High | 0.751 | 0.759 | 0.739 |
| | | SVM | Low | 0.589 | 0.767 | 0.666 |
| | | | High | 0.563 | 0.747 | 0.642 |

Table 18: Cross-validation regression performance over task $re_{11}, re_{12}$ and $re_{13}$ for five models across low- and high-reliability cohorts (divided by median reliability score). Performance metrics include MAE and RMSE. High-reliability cohort rows are highlighted in gray. The best models for the high and low reliability cohorts are denoted with $(*)$ and $(\star)$, respectively.

| ID | Target | Model | Cohort | Metrics | |
|---|---|---|---|---|---|
| | | | | MAE | RMSE |
| $re_{11}$ | Depression | Linear $(\star, *)$ | Low | 7.052 | 9.646 |
| | | | High | 5.306 | 7.825 |
| | | Lasso | Low | 7.345 | 9.750 |
| | | | High | 5.708 | 7.919 |
| | | Ridge | Low | 9.059 | 9.646 |
| | | | High | 5.320 | 7.826 |
| | | RF | Low | 8.075 | 11.127 |
| | | | High | 6.105 | 8.422 |
| | | SVM | Low | 6.570 | 9.795 |
| | | | High | 5.001 | 8.148 |
| $re_{12}$ | High blood sugar | Linear | Low | 6.560 | 8.783 |
| | | | High | 5.658 | 8.078 |
| | | Lasso | Low | 6.664 | 8.818 |
| | | | High | 5.687 | 8.170 |
| | | Ridge $(\star, *)$ | Low | 6.560 | 8.782 |
| | | | High | 5.656 | 8.077 |
| | | RF | Low | 7.370 | 9.521 |
| | | | High | 6.529 | 9.156 |
| | | SVM | Low | 7.103 | 9.529 |
| | | | High | 6.075 | 9.110 |
| $re_{13}$ | Hearing loss | Linear $(\star, *)$ | Low | 7.847 | 11.477 |
| | | | High | 6.841 | 10.610 |
| | | Lasso | Low | 8.061 | 11.457 |
| | | | High | 7.206 | 10.625 |
| | | Ridge | Low | 7.852 | 11.477 |
| | | | High | 6.849 | 10.610 |
| | | RF | Low | 6.153 | 12.378 |
| | | | High | 6.837 | 11.110 |
| | | SVM | Low | 7.372 | 11.758 |
| | | | High | 5.939 | 10.690 |

## Appendix N. Full predictive results for Bayesian adjustment

Table 19: Cross-validation classification performance over task $cl_{21}$ and $cl_{22}$ without and with Bayesian adjustment, and with regression calibration (RC) using five machine learning models. Metrics include precision, recall, and 95% CIs in brackets. F1 scores are used for determining the best model. The unadjusted results using follow-up data only are highlighted in light gray; Bayesian-adjusted results are highlighted in dark gray. The best models using enrollment data only, using follow-up data only, with Bayesian adjustment, and with RC are denoted with $(*)$, (†), $(\star)$ and (‡), respectively.

| ID | Target | Model | Adjustment | Metrics | | |
|---|---|---|---|---|---|---|
| | | | | Precision | Recall | F1 |
| $cl_{21}$ | Depression | Logistic(†) | No (enrollment) | 0.689 [0.642, 0.709] | 0.687 [0.645, 0.704] | 0.674 |
| | | | No (follow-up) | 0.695 [0.647, 0.709] | 0.693 [0.648, 0.704] | 0.681 |
| | | | RC | 0.689 | 0.687 | 0.674 |
| | | | Yes | 0.730 [0.692, 0.761] | 0.731 [0.687, 0.761] | 0.728 |
| | | RF | No (enrollment) | 0.593 [0.497, 0.671] | 0.591 [0.501, 0.666] | 0.591 |
| | | | No (follow-up) | 0.588 [0.483, 0.671] | 0.588 [0.481. 0.672] | 0.585 |
| | | | RC | 0.584 | 0.585 | 0.583 |
| | | | Yes | 0.655 [0.501, 0.692] | 0.642 [0.493, 0.687] | 0.643 |
| | | DT | No (enrollment) | 0.570 [0.459, 0.665] | 0.564 [0.454, 0.657] | 0.564 |
| | | | No (follow-up) | 0.587 [0.452, 0.651] | 0.588 [0.451, 0.648] | 0.584 |
| | | | RC | 0.561 | 0.561 | 0.560 |
| | | | Yes | 0.582 [0.458, 0.672] | 0.567 [0.448, 0.672] | 0.569 |
| | | GB | No (enrollment) | 0.660 [0.514, 0.698] | 0.663 [0.519, 0.699] | 0.660 |
| | | | No (follow-up) | 0.619 [0.498, 0.680] | 0.624 [0.501, 0.681] | 0.616 |
| | | | RC | 0.605 | 0.612 | 0.599 |
| | | | Yes | 0.679 [0.524, 0.722] | 0.672 [0.522, 0.716] | 0.673 |
| | | SVM $(*, \star, ‡)$ | No (enrollment) | 0.699 [0.619, 0.715] | 0.696 [0.621, 0.713] | 0.683 |
| | | | No (follow-up) | 0.693 [0.614, 0.712] | 0.690 [0.615, 0.707] | 0.676 |
| | | | RC | 0.696 | 0.693 | 0.680 |
| | | | Yes | 0.745 [0.654, 0.783] | 0.746 [0.657, 0.776] | 0.744 |
| $cl_{22}$ | Diabetes | Logistic | No (enrollment) | 0.519 [0.517, 0.774] | 0.720 [0.701, 0.736] | 0.603 |
| | | | No (follow-up) | 0.575 [0.587, 0.804] | 0.722 [0.705, 0.736] | 0.607 |
| | | | RC | 0.591 | 0.720 | 0.603 |
| | | | Yes | 0.521 [0.521, 0.804] | 0.722 [0.713, 0.739] | 0.605 |
| | | RF $(*, \star, ‡)$ | No (enrollment) | 0.650 [0.577, 0.697] | 0.679 [0.608, 0.724] | 0.660 |
| | | | No (follow-up) | 0.649 [0.587, 0.714] | 0.684 [0.622, 0.732] | 0.659 |
| | | | RC | 0.650 | 0.681 | 0.659 |
| | | | Yes | 0.685 [0.568, 0.706] | 0.713 [0.626, 0.730] | 0.692 |
| | | DT | No (enrollment) | 0.629 [0.562, 0.691] | 0.625 [0.541, 0.687] | 0.626 |

| | | | | | | |
|---|---|---|---|---|---|---|
| | | | No (follow-up) | 0.624 [0.567, 0.697] | 0.617 [0.549, 0.696] | 0.619 |
| | | | RC | 0.642 | 0.629 | 0.634 |
| | | | Yes | 0.658 [0.552, 0.593] | 0.661 [0.548, 0.696] | 0.659 |
| | | GB (†) | No (enrollment) | 0.658 [0.580, 0.711] | 0.699 [0.622, 0.734] | 0.664 |
| | | | No (follow-up) | 0.688[0.591, 0.724] | 0.705 [0.634, 0.746] | 0.675 |
| | | | RC | 0.637 | 0.681 | 0.648 |
| | | | Yes | 0.645 [0.568, 0.713] | 0.696 [0.626, 0.739] | 0.655 |
| | | SVM | No (enrollment) | 0.699 [0.619, 0.715] | 0.696 [0.621, 0.713] | 0.603 |
| | | | No (follow-up) | 0.519 [0.515, 0.791] | 0.720 [0.682, 0.737] | 0.603 |
| | | | RC | 0.519 | 0.720 | 0.603 |
| | | | Yes | 0.745 [0.654, 0.783] | 0.746 [0.657, 0.776] | 0.605 |

Table 20: Cross-validation regression performance over task $re_{21}$ and $re_{22}$ without and with Bayesian adjustment, and using RC using five machine learning models. Metrics include MAE and RMSE with 95% CIs in brackets. Results with Bayesian adjustment applied are highlighted in gray. The unadjusted results using follow-up data only are highlighted in light gray; Bayesian-adjusted results are highlighted in dark gray. The best models using enrollment data only, using follow-up data only, with Bayesian adjustment, and using RC are denoted with $(*)$, (†), $(\star)$ and (‡), respectively.

| ID | Target | Model | Adjustment | Metrics | |
|---|---|---|---|---|---|
| | | | | MAE | RMSE |
| $re_{21}$ | Diabetes | Linear $(\star)$ | No (enrollment) | 5.566 [5.297, 5.979] | 7.580 [7.321, 8.122] |
| | | | No (follow-up) | 5.580 [5.367, 5.989] | 7.583 [7.340, 8.083] |
| | | | RC | 6.638 | 9.055 |
| | | | Yes | 4.671 [4.523, 5.199] | 6.475 [6.300, 6.988] |
| | | Lasso $(*)$ | No (enrollment) | 5.648 [5.508, 5.982] | 7.672 [7.414, 8.088] |
| | | | No (follow-up) | 5.644 [5.411, 5.914] | 7.649 [7.402, 7.998] |
| | | | RC | 6.524 | 9.105 |
| | | | Yes | 4.625 [4.554, 4.952] | 6.471 [6.296, 6.975] |
| | | Ridge (†, ‡) | No (enrollment) | 5.564 [5.300, 5.972] | 7.578 [7.330, 8.117] |
| | | | No (follow-up) | 5.578 [5.363, 5.985] | 7.581 [7.340, 8.068] |
| | | | RC | 6.587 | 8.989 |
| | | | Yes | 4.466 [4.521, 5.196] | 6.471 [6.296, 6.975] |
| | | RF | No (enrollment) | 6.349 [5.719, 7.362] | 8.924 [7.752, 10.270] |
| | | | No (follow-up) | 6.172 [5.647, 7.149] | 8.331 [7.789, 9.678] |
| | | | RC | 7.786 | 10.856 |
| | | | Yes | 5.652 [4.883, 6.303] | 7.658 [6.746, 8.510] |
| | | SVM | No (enrollment) | 5.758 [5.514, 6.161] | 8.051 [7.747, 8.600] |
| | | | No (follow-up) | 5.988 [5.731, 6.409] | 8.185 [7.863, 8.786] |
| | | | RC | 7.920 | 10.692 |
| | | | Yes | 4.967 [4.825, 5.484] | 6.831 [6.684, 7.431] |
| $re_{22}$ | High cholesterol | Linear $(*, \star, ‡)$ | No (enrollment) | 5.157 [5.073, 5.273] | 6.901 [6.857, 7.002] |
| | | | No (follow-up) | 5.312 [5.242, 5.420] | 6.985 [6.946, 7.078] |
| | | | RC | 5.151 | 6.865 |
| | | | Yes | 4.946 [4.892, 5.217] | 6.357 [6.277, 6.606] |
| | | Lasso | No (enrollment) | 5.360 [5.245, 5.494] | 7.055 [6.974, 7.176] |
| | | | No (follow-up) | 5.505 [5.406,5.617] | 7.157 [7.074, 7.261] |
| | | | RC | 5.342 | 7.025 |
| | | | Yes | 5.205 [5.128, 5.471] | 6.603 [6.504, 6.871] |
| | | Ridge $(\star, †, ‡)$ | No (enrollment) | 5.157 [5.074, 5.278] | 6.901 [6.857, 7.001] |
| | | | No (follow-up) | 5.312 [5.240, 5.419] | 5.985 [6.946, 7.080] |
| | | | RC | 5.151 | 6.865 |
| | | | Yes | 4.947 [4.892, 5.217] | 6.357 [6.277, 6.607] |

| | | | | | |
|---|---|---|---|---|---|
| | | RF | No (enrollment) | 5.790 [5.606, 6.298] | 7.806 [7.668, 8.539] |
| | | | No (follow-up) | 5.871 [5.730, 6.430] | 7.780 [7.664, 8.546] |
| | | | RC | 5.574 | 7.559 |
| | | | Yes | 5.246 [5.210, 5.964] | 7.041 [6.838, 7.765] |
| | | SVM | No (enrollment) | 4.759 [4.691, 4.942] | 7.068 [6.974, 7.238] |
| | | | No (follow-up) | 5.227 [5.148, 5.407] | 7.777 [7.074, 7.378] |
| | | | RC | 4.934 | 6.987 |
| | | | Yes | 4.786 [4.836, 5.289] | 6.561 [6.424, 6.873] |

## Appendix O. Sensitivity analysis of variable normalization for reliability scores

To evaluate whether variables with naturally larger onset age variance disproportionately dominate the raw participant-level reliability score ($r_i = \sum_{j=1}^{p} |\tilde{D}_{ij}|$), we conducted a comparative sensitivity analysis. We implemented a column-wise normalized variant of our pipeline where the absolute age differences in the completed discrepancy matrix $\tilde{D}$ are rescaled condition-by-condition to a bounded range on $[0, 1]$ prior to summing:

$$\tilde{D}_{ij}^{\text{norm}} = \frac{|\tilde{D}_{ij}| - \min_k |\tilde{D}_{kj}|}{\max_k |\tilde{D}_{kj}| - \min_k |\tilde{D}_{kj}|} \tag{4}$$

We then aggregated these normalized discrepancies to construct an alternative set of participant-level reliability scores:

$$r_i^{\text{norm}} = \sum_{j=1}^{p} \tilde{D}_{ij}^{\text{norm}} \tag{5}$$

The empirical results demonstrate that participant rankings are highly robust to variable-level variance discrepancies:

- **Rank Correlation:** The Spearman rank correlation ($\rho$) between the raw-sum reliability scores ($r_i$) and the normalized reliability scores ($r_i^{\text{norm}}$) is **0.976** across the full cohort ($n = 25,018$). Across all task-specific sub-cohorts ($cl_{11}$–$cl_{13}$ and $re_{11}$–$re_{13}$), the correlation remains tightly bounded between 0.93 and 0.97.
- **Cohort Assignment Agreement:** When dividing participants into high- and low-reliability groups using a median split, the cohort assignment agreement between the two scoring methods is extremely high, ranging from **91.5%** to **95.5%** across all downstream classification and regression tasks.

Because of this exceptionally high overlap, downstream predictive performance metrics (precision, recall, MAE, and RMSE) remain largely unchanged. We retain the unscaled sum-of-differences formulation in our main pipeline for its simplicity and direct interpretability in years of reporting discrepancy, but highlight normalized scoring as a simple, robust modification for future applications to cohorts with wider variance imbalances.

## Appendix P. Summary of Inconsistency Types in CanPath data

The following inconsistency types were identified across 97,408 participants with both enrollment and follow-up data. Percentages indicate the proportion of participants exhibiting at least one inconsistency of each type.

**Onset age inconsistency (57.1% of participants).** This inconsistency occurs when the age at which a participant first reported being diagnosed with a condition differs between enrollment and follow-up surveys. For example, a participant may report being diagnosed with diabetes at age 45 during enrollment, but report age 50 at follow-up. Since the age of first diagnosis is a fixed historical event, any difference between timepoints is considered inconsistent.

**Ever-had inconsistency (55.1% of participants).** This inconsistency occurs when a participant reports "yes" to having a condition at enrollment, but "no" at follow-up (i.e., a $1 \rightarrow 0$ transition). For example, a participant may report "yes" to ever having asthma at enrollment, but "no" at follow-up. Such transitions are inconsistent because one cannot "un-have" a previously reported condition. Note that $0 \rightarrow 1$ transitions are acceptable, as conditions can develop over time.

**Calculated age vs date inconsistency (0.24% of participants).** This inconsistency occurs when the difference in calculated age between enrollment and follow-up does not match the time gap between the two questionnaire completion dates. For example, if questionnaires are completed 5 years apart, but reported ages only differ by 2 years, this is flagged as inconsistent. A tolerance of $\pm 2$ years is applied to account for integer rounding in age reporting.