Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Accounting for latent class uncertainty in trajectory-specific risk estimation: the Targeted Compositional Corner Estimator Trajectory-specific risk estimation under latent class uncertainty

Pinyan Liu
Duke-NUS Medical School
Murthy Mittinty
Flinders University
Bibhas Chakraborty
Duke-NUS Medical School
Marco A. Peres
marco.peres@duke-nus.edu.sg

Duke-NUS Medical School

Research Article

Keywords: group-based trajectory modelling, causal inference, compositional data, latent class, periodontitis, diabetes

Posted Date: November 16th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11276349/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Accounting for latent class uncertainty in trajectory-specific risk estimation: the Targeted Compositional Corner Estimator

Trajectory-specific risk estimation under latent class uncertainty

Pinyan Liu, PhD[1]; Murthy Mittinty, PhD[2]; Bibhas Chakraborty, PhD[1,5,6]; Marco A. Peres, PhD[3,4]*

[1] Centre for Biomedical Data Science, Duke-NUS Medical School, Singapore, Singapore 169857

[2] College of Medicine and Public Health, Flinders University, Adelaide, Australia 5042

[3] Programme in Health Services Research and Population Health, Duke-NUS Medical School, Singapore, Singapore 169857

[4] National Dental Research Institute Singapore, National Dental Centre Singapore, Singapore, Singapore 169857

[5] Department of Statistics and Data Science, National University of Singapore, Singapore Singapore

[6] Department of Biostatistics and Bioinformatics, Duke University, Durham, NC, USA

*Corresponding author: Marco A. Peres, National Dental Centre Singapore

30 Hospital Boulevard, Level 22, Singapore, Singapore 168583. Email: marco.peres@duke-nus.edu.sg

Word count:

Abstract: 277

Main text: 4634

Figures: 3

Tables: 6

Keywords: group-based trajectory modelling; causal inference; compositional data; latent class; periodontitis; diabetes.

Abstract

Background: In many chronic diseases, risk depends on how a biomarker evolves over time rather than on its value at a single visit. These longitudinal patterns are commonly summarized as latent trajectory groups, but the uncertainty in class membership is usually ignored by assigning each individual to the single most likely group, which can bias trajectory-specific risk estimates, especially for small groups.

Methods: We developed the Targeted Compositional Corner Estimator (TCCE), a method that retains the full posterior class-membership probabilities and estimates trajectory-specific counterfactual risks using compositional data methods, mapping the probabilities to an unconstrained space with the isometric log-ratio transformation and evaluating risk at the corners representing pure membership. We compared the TCCE with hard assignment, probability-weighted averaging, and targeted maximum likelihood estimation with hard labels in simulations under well-classified, misclassification-stress, and nonlinear-interaction scenarios (200 replicates; n = 2000), including additional conditions with correlated covariates and small, unbalanced classes. We then applied all methods to linked clinical and administrative records from 1,614 Singaporean adults with diabetes, estimating the association between three-year HbA1c trajectories and severe periodontitis, with reduced dentition as a secondary outcome.

Results: Across all simulation scenarios, the TCCE achieved the lowest overall root mean squared error, with the greatest gains under nonlinear interaction, and its advantage persisted under correlated covariates and smaller, unbalanced samples. The point estimates were invariant to the corner offset. In the application, the high-increasing HbA1c trajectory carried the highest risk of severe periodontitis, and the TCCE produced the narrowest compatibility intervals.

Conclusions: The TCCE provides a practical approach for estimating trajectory-specific risks while accounting for latent class uncertainty, and it applies more broadly to exposures represented as probabilities over categories.

1. Background

In many chronic diseases, management depends on how a biomarker behaves over years rather than on its value at a single visit, and two patients with the same current reading may have reached it along very different paths. Whether these long-term patterns, rather than the level at one moment, carry different risks of later harm is often the question of real interest (1). Answering it is a causal problem, and a hard one, because the patterns are not observed directly; they are recovered from the data as latent trajectory groups, estimated with uncertainty. Diabetes is a clear example. Two people can present with the same glycated hemoglobin (HbA1c) today, yet have arrived there along different paths, one holding steady for years and the other rising, and whether those paths carry different risks of periodontal disease is something a single reading cannot tell us (2,3). To study questions of this kind, epidemiologists often summarize repeated measurements into a small number of trajectory groups using group-based or latent class trajectory models (4–6), Then ask how each group relates to a later outcome. Similar summaries are used in psychiatry (7), cardiovascular research (8), and developmental studies (9).

The causal summaries used most often in epidemiology, such as the average treatment effect and the average effect of treatment on the treated (10,11), are defined at the individual level. These are not always the quantities of interest. When the exposure is a long-term pattern rather than a point intervention, the useful question is the risk associated with each pattern, that is, the risk under each trajectory group. Estimating such quantities raises two difficulties that ordinary causal analysis does not face. First, trajectory groups are not observed; they are latent variables, recovered from the data with uncertainty.

Second, causal analysis built on latent trajectory groups is uncommon, and the tools that do exist handle the uncertainty in group membership poorly.

Standard practice assigns each person to the most probable class and treats that assignment as known. This hard plug-in approach ignores classification error, which can bias risk estimates and inflate variability, particularly for small or poorly separated classes (12–15). A softer alternative uses the posterior class probabilities as weights when averaging the outcome; this keeps some of the uncertainty but estimates a weighted average over observed memberships rather than the risk under pure membership, and it offers no formal efficiency guarantee (13,15). Targeted minimum loss-based estimation applied to the hard labels (TMLE-SL) provides doubly robust inference (16,17), yet it still collapses the full posterior into a single assignment.

To address this problem, we developed the Targeted Compositional Corner Estimator (TCCE). The method was built for the specific structure of the problem: posterior class probabilities are compositional, in that they are positive and sum to one, and the quantity of interest is the risk at pure class membership. The TCCE keeps the full posterior, maps it from the probability simplex into unconstrained Euclidean coordinates using the isometric log-ratio transformation (18,19), fits an outcome regression in that space, and evaluates the fitted surface at the simplex corners that represent pure class membership. A bias-correction step and a tunable denominator stabilize the estimates when a class is small.

This paper proceeds in three steps. We first describe and compare four estimators: the hard plug-in, the soft plug-in, TMLE-SL, and the TCCE. We then assess their bias and root mean squared error in a simulation study under three scenarios, namely well classified, misclassification stress, and nonlinear class by covariate interaction, using 200 replicates

of n = 2000. Guided by the simulation results, we apply the estimators to linked clinical and administrative data from a diabetes cohort in Singapore (20), estimating trajectory-specific risks of severe periodontitis and, in a subsample, reduced dentition.

2. Methods

2.1 Notation, estimands, and identification

Let $i = 1, \dots, n$ index subjects. For each person, we observe baseline covariates W_i, a binary outcome Y_i, and an $n \times K$ matrix P of posterior class probabilities from a fitted trajectory model, where P_{ik} is the estimated posterior probability that the person i belongs to the trajectory class k. These posterior probabilities are produced by the trajectory model. They are not weights used in fitting that model; they are inputs to the downstream risk estimators described below.

The target estimands are the class-specific counterfactual risks

$$\psi_k = E\{Y(C = k)\}, \quad k = 1, \dots, K,$$

where $Y(C = k)$ is the outcome that would be observed under membership in class k. With $K = 3$ we report three class-specific risks (ψ_1, ψ_2, ψ_3) and four contrasts. The risk differences are

$$RD_{21} = \psi_2 - \psi_1, \quad RD_{31} = \psi_3 - \psi_1,$$

and the risk ratios are

$$RR_{21} = \psi_2/\psi_1, \quad RR_{31} = \psi_3/\psi_1.$$

All contrasts use the Low-stable class as the reference.

Identification of ψ_k relies on the following four assumptions.

1. Consistency: if $C_i = k$ then $Y_i = Y_i(C = k)$.
2. No interference: one person's class does not affect another person's outcome.
3. Conditional exchangeability:

$$Y_i(C = k) \perp C_i \mid W_{0i}, \quad k = 1, \ldots, K,$$

where W_0 is the baseline confounder set.

4. Positivity: $\Pr(C_i = k \mid W_{0i} = w) > 0$ for every class k and every covariate value w with positive density (21).

Because ψ_k is a total effect, variables that follow the trajectory, such as final HbA1c and cumulative adherence, are mediators and are excluded from W_0. This exchangeability condition is stated at baseline; it does not by itself rule out time-varying common causes that arise during the exposure window and affect both the trajectory and the outcome. We therefore interpret ψ_k as the total effect of a trajectory pattern under baseline exchangeability and return to this point in the Discussion. Further detail is given in Supplementary Material S1.

2.2 Estimators

Hard plug-in. Each person is assigned to the class with the highest posterior probability, $\hat{C}_i = \text{argmax}_k P_{ik}$, and ψ_k is estimated by the sample mean of Y_i among people assigned to class k. The approach is simple but discards classification uncertainty.

Soft plug-in. The posterior probabilities serve as weights,

$$\hat{\psi}_k = \frac{\sum_i P_{ik}\, Y_i}{\sum_i P_{ik}}.$$

This retains some uncertainty, but it estimates a weighted average over the observed posterior values rather than the risk under pure class membership, and it offers no formal efficiency guarantee.

Multivariable TMLE with hard labels (TMLE-SL). The hard label $\hat{C}_i \in \{1,2,3\}$ is treated as a multivalued exposure. A multinomial logistic model estimates the generalised propensity score $g_k(W_i) = \Pr(\hat{C}_i = k \mid W_i)$, with estimates truncated to $[\delta_l, 1-\delta_l]$, $\delta_l = 0.01$, and renormalised. A logistic outcome model

$$\text{logit}Q(A_i, W_i) = \alpha_0 + \alpha_A A_i + \alpha_W^{\top} W_i$$

gives initial predictions $Q_k^0(W_i)$ for each class, truncated to $[\delta_Q, 1-\delta_Q]$, $\delta_Q = 10^{-6}$. The targeting step uses the clever covariates $H_{ki} = \mathbb{1}(\hat{C}_i = k)/g_k(W_i)$ in a logistic fluctuation, giving updated predictions $Q_k^*(W_i)$. The estimate is $\hat{\psi}_k^{\text{TMLE-SL}} = n^{-1}\sum_i Q_k^*(W_i)$. Standard errors come from the efficient influence function

$$IF_{ik} = \frac{\mathbb{1}(\hat{C}_i = k)}{g_k(W_i)}\{Y_i - Q^*(\hat{C}_i, W_i)\} + Q_k^*(W_i) - \hat{\psi}_k,$$

with $\widehat{SE}_k = n^{-1/2}\,\text{SD}(IF_{ik})$. For $K = 3$ we implemented a multivariate version that jointly targets the four pairwise contrasts (16,17). TMLE-SL retains standard semiparametric theory but still reduces the posterior to a single label. Full specification is given in Supplementary Material S2.

Targeted Compositional Corner Estimator (TCCE). The TCCE avoids hard classification. The posterior vector $p_i = (p_{i1}, \dots, p_{iK})$ lies on the $(K-1)$-dimensional probability simplex, so all entries are positive and sum to one. After a small offset $\epsilon = 10^{-8}$ and renormalisation,

$$\tilde{p}_{ik} = \frac{p_{ik} + \epsilon}{\sum_{j=1}^{K}(p_{ij} + \epsilon)},$$

the isometric log-ratio (ILR) transformation (18) maps each composition to $K-1$ unconstrained coordinates, $z_i = \mathrm{ilr}(\tilde{p}_i) \in \mathbb{R}^{K-1}$; for $K = 3$ this gives $z_i = (z_{i1}, z_{i2})$. Pure class membership corresponds to the simplex corners

$$e_k(\epsilon') = (\epsilon', \dots, 1-(K-1)\epsilon', \dots, \epsilon'), \quad \epsilon' = 0.03,$$

with the large value $1-(K-1)\epsilon'$ at position k. These maps to the ILR corner points $z_k^* = \mathrm{ilr}\{e_k(\epsilon')\}$.

A logistic outcome regression is fitted using the baseline confounders and the ILR coordinates,

$$\mathrm{logit}Q(W_i, z_i) = \gamma_0 + \gamma_W^{\mathsf{T}} W_i + \gamma_1 z_{i1} + \gamma_2 z_{i2} + \gamma_{11} z_{i1}^2 + \gamma_{22} z_{i2}^2.$$

The risk at corner k is obtained by substituting z_k^* while keeping each person's own covariates,

$$Q_k^0(W_i) = \mathrm{expit}\{\widehat{\mathrm{logit}Q}(W_i, z_k^*)\},$$

truncated to $[0.001, 0.999]$. Quadratic ILR terms were used in the simulation; the application used linear ILR terms only to avoid extreme extrapolation at the corners. Because the isometric log-ratio coordinates of a corner are identical across subjects,

moving the corner shifts the fitted linear predictor by a subject-independent constant, which the one-parameter logistic fluctuation described below absorbs exactly. The point estimates and their standard errors are therefore invariant to the offset ϵ'; we confirmed this numerically across ϵ' from 0.001 to 0.20 and across shrinkage values (Supplementary Table S4).

The posterior probability for class k provides a weight $w_{ik} = \max(P_{ik}, 0.01)$, with raw denominator $D_k^{\text{raw}} = \sum_i w_{ik}$. For small classes, this denominator is unstable, so it is shrunk toward a balanced target,

$$D_k(\lambda) = \lambda \frac{n}{K} + (1 - \lambda) D_k^{\text{raw}},$$

where K is the number of classes, so the balanced target is n/K (here $n/3$), and $\lambda \in [0,1]$ controls the amount of shrinkage. When $\lambda = 0$ no shrinkage is applied; when $\lambda = 1$ every class shares the same effective denominator. Scaled weights $H_{ik}^{\text{raw}} = n w_{ik} / D_k(\lambda)$ are truncated at their 95th percentile to prevent a few large weights from dominating. A one-dimensional logistic fluctuation then estimates a scalar ϵ_k by weighted quasi-binomial regression, giving the bias-corrected prediction

$$Q_k^*(W_i) = \text{expit}\{\text{logit} Q_k^0(W_i) + \epsilon_k\},$$

which makes the posterior-weighted residual approximately zero. This step resembles targeting in TMLE but solves a different estimating equation. The class-specific risk estimate is

$$\hat{\psi}_k^{\text{TCCE}} = \frac{\sum_i w_{ik}\, Q_k^*(W_i)}{D_k(\lambda)}.$$

An approximate sandwich-type standard error uses the linearized contribution

$$V_{ik} = H_{ik}\{Y_i - Q_k^*(W_i)\} + H_{ik}\{Q_k^*(W_i) - \hat{\psi}_k\}, \quad \widehat{SE}_k = n^{-1/2}\,\mathrm{SD}(V_{ik}).$$

This is an approximation rather than the efficient influence function of a semiparametric model. TCCE uses no propensity score and makes no double-robustness claim. The ILR transformation, corner construction, fluctuation step, and variance derivation are set out in full in Supplementary Material S3 to S5.

A summary of how the four estimators handle trajectory class uncertainty is presented in Table 1.

2.3 Choosing the shrinkage parameter

In the simulation, where the truth is known, λ was chosen to minimise a multi-objective loss across the seven primary estimands. For a candidate λ, let $r_j(\lambda)$ be the RMSE for estimand j, with risk ratios on the log scale. The loss standardises each component by the corresponding soft plug-in RMSE,

$$L(\lambda) = \sum_{j=1}^{7} \left\{ \frac{r_j(\lambda)}{r_j(\mathrm{Soft})} \right\}^2,$$

so, values below seven indicate that the TCCE beats the soft plug-in on average, the simplest method that uses the full posterior. The 200 replicates were split in half into training and test sets 20 times, and the selected value was the mode of the chosen values over the grid $\{0, 0.01, 0.02, 0.05, 0.10\}$.

In the real-data application, where no ground truth exists, λ was chosen by a bias-penalised bootstrap criterion. For each candidate, λ, 500 bootstrap replicates gave a distribution of

risk contrasts, and the criterion added to the bootstrap variance a penalty equal to the deviation of the mean contrast from the unshrunk reference at $\lambda = 0$. In the periodontitis analysis, the loss increased monotonically across the grid, and $\lambda = 0$ was selected. Details are given in Supplementary Material S6.

2.4 Simulation study

We simulated 200 datasets of n = 2000 subjects with T = 6 HbA1c measurements each. Baseline covariates were generated independently: age $\sim \text{Uniform}(30,80)$, male $\sim \text{Bernoulli}(0.5)$, minority $\sim \text{Bernoulli}(0.35)$, income low $\sim \text{Bernoulli}(0.3)$, smoking history $\sim \text{Bernoulli}(0.4)$, and diabetes category $\sim \text{Multinomial}(0.55,0.35,0.10)$ for none, prediabetes, and diabetes. True class membership ($K = 3$) was drawn from a multinomial logistic model with class 1 as the reference,

$$\Pr(C_i = k \mid W_i) = \frac{\exp(\eta_{ik})}{\sum_{j=1}^{3} \exp(\eta_{ij})}, \quad \eta_{i1} = 0,$$

where the linear predictors η_{ik} depend on the covariates through coefficient vectors β_2 and β_3 (full values in Supplementary Material S7). HbA1c trajectories followed class-specific intercepts $\alpha_c \in \{5.3,6.3,7.7\}$ and slopes $s_c \in \{0.00,0.06,0.14\}$, with subject-level random deviations and measurement noise.

Three scenarios were examined. In the well-classified scenario the trajectories matched a quadratic model fitted by the flexmix package in R (22), so the classes were well separated. In the misclassification stress scenario, class-specific cubic and sinusoidal terms

$$f_c(t) = \gamma_{c,\text{cubic}}\, \tilde{t}^{\,3} + \gamma_{c,\text{sine}} \sin(\pi \tilde{t}), \quad \tilde{t} = \frac{t - (T+1)/2}{T/2},$$

heteroskedastic noise, and heavy-tailed Student's-$t(4)$ errors were added, making the estimation model misspecified. In the nonlinear interaction scenario, the stress was retained and the outcome model also included class by age and class by adherence interactions,

$$\beta_{\text{class}\times\text{age}} = (0.0, 0.15, 0.4) \text{ per decade}, \quad \beta_{\text{class}\times\text{adherence}} = (0.0, 0.1, 0.3),$$

producing a nonlinear response surface on the simplex. This scenario separates estimators that average over observed posteriors from estimators that extrapolate to the corners. The binary outcome was generated from a logistic model with a class effect $\beta_{c,\text{class}} \in \{-0.7, -0.2, 0.3\}$ and a diabetes-category effect $\beta_{0d} \in \{-1.5, -0.3, 0.0\}$ (full outcome model in Supplementary Material S7). True class-specific risks were computed by Monte Carlo integration with 200000 draws.

To address the concern that independently generated covariates and a large, balanced sample understate sampling variability, we repeated the nonlinear-interaction scenario under two additional conditions: baseline covariates made mutually dependent through a Gaussian copula, and a smaller sample ($n = 600$) with a small, unbalanced High-increasing class (about 10%, matching the application). We also assessed the sensitivity of the corner estimator to class separation, by varying the within-class trajectory noise at a fixed true risk, and to the corner offset ϵ', over the grid $\{0.001, 0.005, 0.01, 0.02, 0.03, 0.05, 0.08, 0.12, 0.20\}$. Full specifications and results are given in Supplementary Tables S4 to S6.

2.5 Real-world application

Study population. We used linked data from the Singapore Diabetes Registry and the Outpatient Dental Record (20). For each patient, the first derivation-eligible periodontal

visit was the index date, and the exposure window was three years before it. Patients needed at least two HbA1c measurements spanning at least 0.5 years within this window (23), giving 1614 eligible subjects (Figure 1).

Outcomes. Periodontal status was classified using the CDC/AAP case definitions (24). Because periodontitis prevalence was 98.3%, the primary outcome was severe periodontitis with a prevalence of 74% (coded 1 = severe; 0 = none, mild, or moderate). The secondary outcome, reduced dentition, was defined as fewer than 20 teeth present at the index visit (coded 1 = unfavorable; 0 = favorable).

Trajectory modelling. HbA1c trajectories were modelled with a latent class linear mixed model in continuous time (lcmm package (25)). Models with K = 2, 3, and 4 classes were fitted; BIC supported K = 3, and K = 4 was not retained because one class held only 28 patients (1.7%). The fitted model produced the posterior probability matrix P used by all four estimators.

Confounders. The adjustment set was age, sex, ethnicity, housing type, diabetes duration, body mass index (BMI), and cardiovascular disease history. Baseline diabetes category was excluded because it lies on the causal pathway between HbA1c trajectory and periodontitis.

Missing data. Confounder missingness ranged from 0.7% to 27.5%. We used multiple imputation by chained equations with predictive mean matching (M = 10; mice package (26)). The outcomes (periodontitis status, tooth count) were complete and were not imputed, but they were included as auxiliary variables to support the missing-at-random assumption (27). Because the HbA1c trajectory data were complete, the posterior matrix

286 P was identical across imputations, so only the confounder-dependent estimators were re-
287 run on each dataset. Results were combined by Rubin's rules (28): the pooled estimate is
288 the mean across imputations, and the total variance combines the within-imputation
289 variance from B = 500 bootstrap replicates with the between-imputation variance. The
290 fraction of missing information (FMI) summarizes the contribution of imputation
291 uncertainty. Multiple imputation draws from a Bayesian predictive distribution, but the
292 combining rules give frequentist-valid intervals (29).

293 **Secondary analysis: reduced dentition with baseline adjustment.** The primary analysis
294 lacked baseline periodontal status. Among the 1614 patients, 572 (35.4%) had at least one
295 earlier dental visit with a recorded tooth count. For these patients, we took the baseline
296 tooth count from the earliest prior visit and added it as a continuous confounder, which
297 allows a check of whether glycaemic trajectory predicts tooth loss beyond pre-existing
298 dental status.

299 **Inference.** Point estimates and bootstrap standard errors (B = 500) were obtained from the
300 first imputed dataset; all four estimators were then re-run on each of the 10 imputed
301 datasets, and the results were pooled. We report 95% compatibility intervals from the
302 pooled standard errors (30), and we use compatibility language rather than significance
303 language throughout.

304 **3. Results**

305 Figure 2 shows that TCCE has the lowest overall RMSE in all three scenarios. The
306 advantage is consistent and grows as conditions become harder, from a small edge over the
307 soft plug-in in the well-classified scenario to a clear separation under nonlinear interaction.

3.1 Simulation: well classified

Table 2 (Panel A) shows bias and RMSE for the four primary contrast estimands. TCCE had the lowest RMSE for six of seven estimands: both risk differences (0.030 for RD_{21}; 0.021 for RD_{31}) and RR_{21} (0.11). For RR_{31} the soft plug-in was slightly better (0.19 versus 0.23), and the hard plug-in performed worst (RR_{31} RMSE 0.47). TMLE-SL had the highest RMSE for the risk differences, reflecting variability from inverse-probability weighting with small classes.

3.2 Simulation: misclassification stress

Under stress (Table 2, Panel B), the TCCE kept the lowest RMSE for both risk differences (0.02 each) and for RR_{21} (0.10). The soft plug-in had a small advantage for RR_{31} (0.17 versus 0.19). The hard plug-in was worse for the risk ratios (0.25 and 0.51), and TMLE-SL showed larger bias for the risk differences than in the well-classified scenario.

3.3 Simulation: nonlinear interaction

Under nonlinear interaction (Table 2, Panel C), the TCCE achieved the lowest RMSE for all four contrasts. The mean RMSE across the seven estimands was 0.056 for the TCCE, 0.073 for the soft plug-in, 0.13 for the hard plug-in, and 0.14 for TMLE-SL. The mean contrast RMSE was 28% lower for the TCCE than for the soft plug-in (0.076 versus 0.11).

Across all three scenarios, the weighting-based estimators (soft plug-in, hard plug-in, and TCCE) showed a small negative bias for the class-specific risks, meaning they slightly underestimated absolute risk (Supplementary Tables S1 to S3). Because this bias was close to uniform across classes, it largely cancelled in the contrasts: for the High-increasing

versus Low-stable difference, the bias of the TCCE stayed between 0.000 and 0.013 in absolute value, whereas TMLE-SL, which overestimated the reference class and underestimated the high-risk class, carried a contrast bias of 0.056 to 0.069. This is why the TCCE reaches a low contrast RMSE despite a modest per-class bias.

3.4 Robustness and sensitivity analyses

The corner estimator's point estimates were invariant to the offset ϵ': across ϵ' from 0.001 to 0.20 and across shrinkage values from 0 to 0.3, the class-specific risks were identical to machine precision, with a range of order 10^{-16} (Supplementary Table S4). Under the harder data-generating model, correlated covariates together with a smaller, unbalanced sample (High-increasing class about 11%) roughly doubled the root mean squared error of the High-increasing versus Low-stable contrast for both the TCCE and the soft estimator, while preserving their ordering (Supplementary Table S6). Varying class separation at a fixed true risk showed that the small downward bias of the TCCE for the extreme class grew as the classes became harder to identify, from about 0.007 when they were well identified to about 0.050 when they were poorly identified, and it cancelled substantially in the contrasts (Supplementary Table S5).

3.5 Application: trajectory classes

Three trajectory classes were identified (Figure 3): Low-stable (n = 1343, 83.2%; mean HbA1c 7.0%), Moderate (n = 115, 7.1%; mean HbA1c 8.8%), and High-increasing (n = 156, 9.7%; mean HbA1c 9.5%). For reference, HbA1c below 5.7% is considered normal, 5.7% to 6.4% indicates prediabetes, and 6.5% or above indicates diabetes, with clinical guidelines generally recommending a treatment target below 7.0% for most adults with

diabetes (31). The Low-stable class therefore sits near the target, while the Moderate and High-increasing classes reflect progressively poorer control. Classification quality was good overall (mean maximum posterior 0.95; 91.0% well classified), with more overlap for the Moderate and High-increasing classes (mean maximum posterior 0.87 and 0.88; 70.4% and 73.1% well classified). Baseline characteristics appear in Table 3.

3.6 Application: severe periodontitis

All four estimators identified the High-increasing trajectory as carrying the highest risk of severe periodontitis (point estimates 0.75 to 0.78; Table 5). For the High-increasing versus Low-stable contrast, every estimator gave a positive risk difference (0.014 to 0.048) and a risk ratio above one (1.018 to 1.065), and all compatibility intervals included the null. The TCCE produced the narrowest interval. The bias-penalized bootstrap selected $\lambda = 0$, so no denominator shrinkage was applied. The Moderate versus Low-stable contrast was near zero across all methods. The FMI was below 0.1% for all estimates.

3.7 Application: reduced dentition

Among the 572 patients with a baseline tooth count, reduced dentition prevalence was 31.5% (Table 4). After adjusting for baseline tooth count, trajectory-specific risks were similar across classes (0.28 to 0.32 under the TCCE; Table 6), and no consistent dose-response pattern emerged. TCCE again gave the narrowest contrast interval for the High-increasing versus Low-stable comparison.

4. Discussion

Trajectory analyses have become ubiquitous across epidemiology, psychiatry, cardiovascular research and developmental science, yet virtually all downstream analyses ignore uncertainty in latent class assignment. This work addresses that fundamental limitation by introducing an estimator that directly targets counterfactual trajectory-specific risks while propagating latent class uncertainty throughout estimation.

We proposed TCCE for estimating trajectory-specific disease risk when class membership is latent and estimated with uncertainty. Rather than reducing the posterior membership vector to a single assigned class, TCCE retains the full posterior and targets the counterfactual risk under pure class membership. In the simulation studies, TCCE achieved the lowest overall RMSE across all three scenarios, with its advantage most pronounced when the outcome interacted nonlinearly with class. In the application, it yielded the narrowest compatibility intervals among the four estimators considered. To our knowledge, this is among the first studies to integrate group-based trajectory modelling of glycaemic control with causal estimation of trajectory-specific periodontal risk.

The estimation problem addressed here is a specific instance of a general one: inference on the effect of an exposure that is not observed directly but inferred as a latent category. In latent class analysis, relating class membership to a distal outcome by assigning each unit to its modal class and treating that assignment as observed is known to bias the resulting associations, because it disregards classification uncertainty; bias-adjusted three-step procedures were developed in response (12,14,15). TCCE may be viewed as a causal analogue of this. In place of a corrected association, it targets a counterfactual risk defined at the vertices of the membership simplex while propagating the entire posterior through

the estimator. The same structure arises in measurement error, where a modal label is a misclassified exposure and the soft estimator is its posterior-weighted correction. TCCE extends this by modelling the outcome as a smooth function of the posterior and evaluating that function where membership is certain. The requisite geometry is provided by compositional data analysis: posterior vectors are positive and sum to one, and therefore lie on the simplex, and the isometric log-ratio transformation (18,19) maps them to an unconstrained space in which the outcome regression is identifiable and the pure-class corners are well defined. The bias-correction step adapts the targeting principle of targeted learning (16), although the estimating equation it solves is specific to the compositional corner.

This construction confers on TCCE an estimand that the other estimators do not share. The soft estimator returns a posterior-weighted average of risks over units with similar membership profiles, that is, a mixture over impure states, whereas TCCE returns the risk under pure membership, which is the estimand of primary interest. The two coincide when the outcome depends approximately linearly on membership, as in the primary application, and diverge under nonlinearity, where corner evaluation recovers the pure-class effect while interior averaging is attenuated toward the population mean. This distinction accounts for the advantage of TCCE in the interaction scenario. Among the comparators, the hard plug-in discarded classification uncertainty and was the most biased, whereas TMLE-SL retained double robustness but collapsed the posterior to a modal label and generally performed between the two simpler estimators. The efficiency gains of TCCE were concentrated in the contrasts, in which a bias that is approximately common across classes largely cancels.

These results suggest practical guidance for selecting among the estimators. When classes are well separated, posterior memberships are near-degenerate, and the outcome model is approximately linear, the soft estimator is an adequate and parsimonious choice. When formal semiparametric efficiency or double-robustness guarantees are required and a modal assignment is defensible, TMLE-SL is preferable. TCCE is most advantageous in the more challenging and frequently encountered setting in which at least one class is small, separation is imperfect so that posteriors retain substantial uncertainty, and the outcome may interact with class. This is precisely the regime in which modal assignment is least reliable and the pure-class contrast is the target of inference. Because its computational cost exceeds that of the soft estimator only marginally, TCCE also affords protection against undetected nonlinearity.

Although motivated by trajectory classes, TCCE is not specific to them. It is applicable whenever an exposure is represented as a probability vector over categories or, more generally, as a composition. Relevant epidemiological examples include latent classes of multimorbidity or of dietary and behavioral patterns, model-derived disease subtypes, and soft labels produced by probabilistic classifiers, together with intrinsically compositional exposures such as microbiome or immune cell-type proportions and time-use budgets. In each case, rather than hardening the probabilistic exposure into a single category, the outcome is modelled as a function of the full composition and the effect is evaluated at the vertices representing pure exposure states.

In the application, the elevated risk of severe periodontitis in the High-increasing trajectory is consistent with the established association between hyperglycemia and periodontal inflammation (2,3). The inclusion of the null in all compatibility intervals is unsurprising

given the modest effect sizes and the high baseline prevalence of severe periodontitis, approximately 74%, and does not constitute evidence of no effect (32). The near-null Moderate versus Low-stable contrast may reflect closer clinical monitoring of patients with moderately elevated HbA1c. In the secondary analysis, adjustment for baseline tooth count attenuated the trajectory gradient, as anticipated when a baseline measure captures disease accumulated under prior exposure, underscoring the value of baseline outcome information when it is available.

This study has limitations. First, baseline periodontal status was unavailable for the primary outcome; the secondary dentition analysis addresses this, but in a smaller subsample. Second, the class-specific risk estimates from the TCCE carry a small downward bias for the most extreme class. This bias does not arise from the corner offset ϵ': because a corner's isometric log-ratio coordinates are identical across subjects, the fluctuation absorbs the offset exactly and the estimates are invariant to it (Supplementary Table S4). The bias instead reflects attenuation of the estimated exposure, as the outcome model is fitted against posterior-derived coordinates that are a noisy proxy for true membership; at a fixed true risk it grew from about 0.007 when classes were well identified to about 0.050 when they were poorly identified (Supplementary Table S5), and it cancelled substantially in the contrasts, which are our primary estimands. Third, identification relies on baseline conditional exchangeability; because the exposure is a three-year trajectory, time-varying common causes that arise within the window and affect both the trajectory and periodontal risk are not addressed by baseline adjustment, and a full treatment would require g-methods and a different estimand, which we leave to future work. Fourth, the TCCE depends on the outcome regression being approximately correct and on a tuned shrinkage parameter, and

it does not provide the double-robustness guarantee of TMLE-SL, so under outcome-model misspecification the bias-correction step may not fully remove confounding. Fifth, the method does not yet rest on a complete semiparametric theory: there is no efficient influence function or natural propensity score for a simplex corner, the variance estimator is an approximation, and inference relies on the bootstrap, which can undercover when a class is very small. Future work should develop theory for compositional-label estimators and test whether more robust versions can be built. The method should also be evaluated beyond glycaemic trajectories and periodontitis, for example, on symptom trajectories, cardiovascular risk patterns, and medication adherence. Larger multi-centre studies are needed to assess performance when trajectory classes are better separated, and minority classes are larger. Natural methodological extensions include continuous outcomes, through a linear outcome model in place of the logistic one, and time-to-event outcomes, where the estimand becomes a trajectory-specific survival or cumulative-incidence curve.

5. Conclusions

TCCE provides a practical way to estimate trajectory-specific outcome risks while accounting for uncertainty in class membership. This matters in clinical epidemiology because patient trajectories are widely used to describe long-term patterns, yet assigning each person to a single class can bias risk estimates, especially for smaller groups. In the glycaemic control and periodontitis application, the method produced narrower compatibility intervals and showed a clear risk pattern across HbA1c trajectories. These findings suggest that TCCE can help researchers use longitudinal clinical data more effectively when linking disease trajectories to later clinical outcomes.

Declarations

Ethics approval and consent to participate. The simulation study involved no human participants. The analysis of de-identified data from the National Dental Centre Singapore Oral Disease Registry was approved by the SingHealth Institutional Review Board (CIRB Reference Code: 2022-2167). The requirement for individual informed consent was waived for this secondary analysis of de-identified registry data. All methods were carried out in accordance with relevant guidelines and regulations.

Consent for publication. Not applicable.

Availability of data and materials. The data underlying this article are available from the National Dental Research Institute Singapore, National Dental Centre Singapore. Restrictions apply; data are available from the authors upon reasonable request with permission of the National Dental Centre Singapore. Reproducibility materials (code) are available upon request.

Competing interests. The authors have nothing to declare.

Clinical trial number. Not applicable.

Funding. Marco A. Peres was supported by the Ministry of Health Singapore through the National Medical Research Council under the Singapore Translational Research Investigator Award (MOH-001416-01). Pinyan Liu received a PhD scholarship from Duke-NUS Medical School. No direct funding was received for this study.

Author contributions. Pinyan Liu: conceptualization, data curation, formal analysis, investigation, methodology, project administration, software, validation, visualization,

writing (original draft). Murthy Mittinty: conceptualization, investigation, methodology, supervision, writing (review and editing). Bibhas Chakraborty: conceptualization, investigation, methodology, writing (review and editing). Marco A. Peres: conceptualization, investigation, methodology, project administration, resources, supervision, writing (review and editing).

Acknowledgements. Not applicable

Use of generative artificial intelligence. During the preparation of this work, the authors used ChatGPT and Claude to check grammar. After using these tools, the authors reviewed and edited the content as needed and took full responsibility for the content of the published article.

References

1. Sanz M, Ceriello A, Buysschaert M, Chapple I, Demmer RT, Graziani F, et al. Scientific evidence on the links between periodontal diseases and diabetes: Consensus report and guidelines of the joint workshop on periodontal diseases and diabetes by the International diabetes Federation and the European Federation of Periodontology. Diabetes Res Clin Pract. 2018 Mar;137:231–41. doi:10.1016/j.diabres.2017.12.001 PubMed PMID: 29208508.

2. Preshaw PM, Alba AL, Herrera D, Jepsen S, Konstantinidis A, Makrilakis K, et al. Periodontitis and diabetes: a two-way relationship. Diabetologia. 2012 Jan;55(1):21–31. doi:10.1007/s00125-011-2342-y PubMed PMID: 22057194; PubMed Central PMCID: PMC3228943.

526 3. Genco RJ, Borgnakke WS. Diabetes as a potential risk for periodontitis: association
527 studies. Periodontol 2000. 2020 Jun;83(1):40–5. doi:10.1111/prd.12270 PubMed
528 PMID: 32385881.

529 4. Group-based multi-trajectory modeling - Daniel S Nagin, Bobby L Jones, Valéria Lima
530 Passos, Richard E Tremblay, 2018 [Internet]. [cited 2024 Apr 26]. Available from:
531 https://journals.sagepub.com/doi/abs/10.1177/0962280216673085

532 5. Nagin DS. Group-based trajectory modeling: an overview. Ann Nutr Metab.
533 2014;65(2–3):205–10. doi:10.1159/000360229 PubMed PMID: 25413659.

534 6. Nagin DS, Odgers CL. Group-based trajectory modeling in clinical research. Annu Rev
535 Clin Psychol. 2010;6:109–38. doi:10.1146/annurev.clinpsy.121208.131413 PubMed
536 PMID: 20192788.

537 7. Musliner KL, Munk-Olsen T, Eaton WW, Zandi PP. Heterogeneity in long-term
538 trajectories of depressive symptoms: Patterns, predictors and outcomes. J Affect
539 Disord. 2016 Mar 1;192:199–211. doi:10.1016/j.jad.2015.12.030 PubMed PMID:
540 26745437; PubMed Central PMCID: PMC4761648.

541 8. Tielemans SMAJ, Geleijnse JM, Menotti A, Boshuizen HC, Soedamah-Muthu SS,
542 Jacobs DR, et al. Ten-year blood pressure trajectories, cardiovascular mortality, and
543 life years lost in 2 extinction cohorts: the Minnesota Business and Professional Men
544 Study and the Zutphen Study. J Am Heart Assoc. 2015 Mar 9;4(3):e001378.
545 doi:10.1161/JAHA.114.001378 PubMed PMID: 25753924; PubMed Central PMCID:
546 PMC4392427.

9. Nagin D, Tremblay RE. Trajectories of boys' physical aggression, opposition, and hyperactivity on the path to physically violent and nonviolent juvenile delinquency. Child Dev. 1999 Oct;70(5):1181–96. doi:10.1111/1467-8624.00086 PubMed PMID: 10546339.

10. Hernan MA, Robins JM. Causal Inference: What If [Internet]. CRC Press; 2025. (Chapman & Hall/CRC Monographs on Statistics & Applied Probab). Available from: https://books.google.com.sg/books?id=_KnHIAAACAAJ

11. Imbens GW, Rubin DB. Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction [Internet]. Cambridge University Press; 2015. Available from: https://books.google.com.sg/books?id=FYeSBwAAQBAJ

12. Nylund-Gibson K, Grimm RP, Masyn KE. Prediction from Latent Classes: A Demonstration of Different Approaches to Include Distal Outcomes in Mixture Models. Structural Equation Modeling: A Multidisciplinary Journal. 2019 Nov 2;26(6):967–85. doi:10.1080/10705511.2019.1590146

13. Loh WW, Kim JS. Evaluating sensitivity to classification uncertainty in latent subgroup effect analyses. BMC Medical Research Methodology. 2022 Sep 24;22(1):247. doi:10.1186/s12874-022-01720-8

14. Vermunt JK. Latent Class Modeling with Covariates: Two Improved Three-Step Approaches. Political Analysis. 2017/01/04 ed. 2017;18(4):450–69. Located at: Cambridge Core. doi:10.1093/pan/mpq025

15. Bakk Z, Oberski DL, Vermunt JK. Relating Latent Class Assignments to External Variables: Standard Errors for Correct Inference. Political Analysis. 2017/01/04 ed. 2017;22(4):520–40. Located at: Cambridge Core. doi:10.1093/pan/mpu003

16. Laan MJ van der, Rose S. Targeted Learning: Causal Inference for Observational and Experimental Data. New York: Springer-Verlag New York Inc.; 2011. 628 p.

17. Schuler MS, Rose S. Targeted Maximum Likelihood Estimation for Causal Inference in Observational Studies. Am J Epidemiol. 2017 Jan 1;185(1):65–73. doi:10.1093/aje/kww165 PubMed PMID: 27941068.

18. Egozcue JJ, Pawlowsky-Glahn V, Mateu-Figueras G, Barceló-Vidal C. Isometric Logratio Transformations for Compositional Data Analysis. Mathematical Geology. 2003 Apr 1;35(3):279–300. doi:10.1023/A:1023818214614

19. Aitchison J. The Statistical Analysis of Compositional Data. Journal of the Royal Statistical Society: Series B (Methodological). 1982;44(2):139–60. doi:https://doi.org/10.1111/j.2517-6161.1982.tb01195.x

20. Marco Aurélio Peres, Pinyan Liu, John Rong Hao Tay. The National Dental Centre Singapore Oral Disease Registry (NDCS-ODR): a high-dimensional longitudinal clinical data resource for epidemiology and inference under structured missingness. International Journal of Epidemiology. 2026;((under review)).

21. Petersen ML, Porter KE, Gruber S, Wang Y, van der Laan MJ. Diagnosing and responding to violations in the positivity assumption. Stat Methods Med Res. 2012

587 Feb;21(1):31–54. doi:10.1177/0962280210386207 PubMed PMID: 21030422;
588 PubMed Central PMCID: PMC4107929.

589 22. Grün B, Leisch F. FlexMix Version 2: Finite Mixtures with Concomitant Variables and
590 Varying and Constant Parameters. J Stat Soft. 2008 Oct 6;28(4):1–35.
591 doi:10.18637/jss.v028.i04

592 23. Nicolaisen SK, le Cessie S, Thomsen RW, Witte DR, Dekkers OM, Sørensen HT, et
593 al. Longitudinal HbA1c patterns before the first treatment of diabetes in routine clinical
594 practice: A latent class trajectory analysis. Diabetes Res Clin Pract. 2024
595 Jun;212:111722. doi:10.1016/j.diabres.2024.111722 PubMed PMID: 38815656.

596 24. Eke PI, Page RC, Wei L, Thornton-Evans G, Genco RJ. Update of the Case Definitions
597 for Population-Based Surveillance of Periodontitis. J Periodontol. 2012
598 Dec;83(12):1449–54. doi:10.1902/jop.2012.110664 PubMed PMID: 22420873;
599 PubMed Central PMCID: PMC6005373.

600 25. Proust-Lima C, Philipps V, Liquet B. Estimation of Extended Mixed Models Using
601 Latent Classes and Latent Processes: The R Package lcmm. J Stat Soft. 2017 Jun
602 1;78(2):1–56. doi:10.18637/jss.v078.i02

603 26. van Buuren S, Groothuis-Oudshoorn K. mice: Multivariate Imputation by Chained
604 Equations in R. J Stat Soft. 2011 Dec 12;45(3):1–67. doi:10.18637/jss.v045.i03

605 27. van Buuren, S. (2012). Flexible Imputation of Missing Data (1st ed.). Chapman and
606 Hall/CRC. https://doi.org/10.1201/b11826.

607 28. Toutenburg H. Rubin, D.B.: Multiple imputation for nonresponse in surveys. Statistical
608 Papers. 1990 Dec 1;31(1):180–180. doi:10.1007/BF02924688

609 29. Rubin DB. Multiple Imputation after 18+ Years. Journal of the American Statistical
610 Association. 1996;91(434):473–89. doi:10.1080/01621459.1996.10476908

611 30. Marshall A, Altman DG, Holder RL, Royston P. Combining estimates of interest in
612 prognostic modelling studies after multiple imputation: current practice and guidelines.
613 BMC Med Res Methodol. 2009 Jul 28;9:57. doi:10.1186/1471-2288-9-57 PubMed
614 PMID: 19638200; PubMed Central PMCID: PMC2727536.

615 31. American Diabetes Association Professional Practice Committee for Diabetes*. 6.
616 Glycemic Goals, Hypoglycemia, and Hyperglycemic Crises: Standards of Care in
617 Diabetes—2026. Diabetes Care. 2025 Dec 8;49(Supplement_1):S132–49.
618 doi:10.2337/dc26-S006

619 32. Greenland S, Mansournia MA, Altman DG. Sparse data bias: a problem hiding in plain
620 sight. BMJ. 2016 Apr 27;352:i1981. doi:10.1136/bmj.i1981 PubMed PMID:
621 27121591.

622

623 **Tables**

624 **Table 1. Comparison of how the four estimators handle trajectory class uncertainty.**

Estimator	Class information used	Main strength	Main caution
Hard plug-in	One most likely class	Easy to explain and reproduce	Ignores class uncertainty
Soft plug-in	Posterior probabilities as weights	Simple way to keep uncertainty	No formal double-robustness guarantee
TMLE-SL	Hard class label as a multivalued exposure	Uses standard TMLE theory for a categorical exposure	Still discards the posterior probability vector
TCCE	Full posterior probability vector	Uses uncertainty and estimates risks at class corners	Variance is approximate; no double-robustness claim

TMLE-SL, targeted minimum loss-based estimation with structured (hard) labels; TCCE, Targeted Compositional Corner Estimator.

Table 2. Bias and RMSE for the four risk contrasts across three simulation scenarios (n = 2000).

Each cell shows bias with RMSE in parentheses; the lowest RMSE in each row is in bold.

Panel A. Well-classified setting

Contrast	True	TCCE	Soft	Hard	TMLE-SL
RD_{21}	0.096	**−0.025 (0.030)**	−0.038 (0.043)	−0.040 (0.057)	−0.062 (0.076)
RD_{31}	0.21	**−0.003 (0.021)**	−0.023 (0.033)	0.006 (0.038)	−0.069 (0.081)
RR_{21}	1.39	**−0.042 (0.11)**	−0.12 (0.17)	−0.078 (0.29)	−0.23 (0.31)
RR_{31}	1.84	0.16 (0.23)	**0.022 (0.19)**	0.26 (0.47)	−0.26 (0.37)

Panel B. Misclassification stress setting

Contrast	True	TCCE	Soft	Hard	TMLE-SL
RD_{21}	0.097	**−0.01 (0.02)**	−0.03 (0.03)	−0.01 (0.04)	−0.05 (0.07)
RD_{31}	0.21	**0.00 (0.02)**	−0.02 (0.03)	0.033 (0.05)	−0.06 (0.07)
RR_{21}	1.39	**0.005 (0.10)**	−0.08 (0.14)	0.039 (0.25)	−0.20 (0.27)
RR_{31}	1.84	0.106 (0.19)	**0.002 (0.17)**	0.35 (0.51)	−0.24 (0.34)

Panel C. Nonlinear interaction setting

Contrast	True	TCCE	Soft	Hard	TMLE-SL
RD_{21}	0.108	**−0.024 (0.024)**	−0.032 (0.039)	−0.043 (0.043)	−0.060 (0.072)

Contrast	True	TCCE	Soft	Hard	TMLE-SL
RD_{31}	0.241	**−0.013 (0.025)**	−0.027 (0.036)	−0.029 (0.048)	−0.068 (0.079)
RR_{21}	1.439	**−0.043 (0.098)**	−0.123 (0.161)	−0.011 (0.231)	−0.245 (0.297)
RR_{31}	1.977	**0.0078 (0.156)**	−0.089 (0.181)	0.283 (0.451)	−0.327 (0.396)

RMSE = root mean squared error; RD = risk difference; RR = risk ratio. Subscripts: 2 = Moderate, 3 = High-increasing, 1 = Low-stable (reference). TCCE = Targeted Compositional Corner Estimator; TMLE-SL = targeted minimum loss-based estimation with structured labels. TCCE rows shown in bold.

Table 3. Baseline characteristics by HbA1c trajectory class: severe periodontitis analysis (n = 1614).

Variable	**Overall**	**Low-stable**	**Moderate**	**High increasing**
N	1614	1343	115	156
Age, years	61.3 (10.7)	62.4 (10.2)	54.8 (11.4)	56.3 (11.4)
Male	940 (58.2%)	787 (58.6%)	60 (52.2%)	93 (59.6%)
Ethnicity				

Chinese	1109 (68.7%)	960 (71.5%)	66 (57.4%)	83 (53.2%)
Indian	242 (15.0%)	189 (14.1%)	17 (14.8%)	36 (23.1%)
Malay	191 (11.8%)	139 (10.3%)	22 (19.1%)	30 (19.2%)
Others	72 (4.5%)	55 (4.1%)	10 (8.7%)	7 (4.5%)
Diabetes duration, years	8.5 (8.0)	8.2 (7.9)	8.7 (7.5)	10.7 (9.1)
BMI, kg/m²	26.8 (5.0)	26.6 (5.0)	27.6 (5.3)	27.9 (5.1)
CVD history	421 (26.1%)	350 (26.1%)	23 (20.0%)	48 (30.8%)
Periodontitis present	1569 (97.2%)	1305 (97.2%)	111 (96.5%)	153 (98.1%)
Severe periodontitis	1195 (74.0%)	986 (73.4%)	87 (75.7%)	122 (78.2%)

Continuous variables: mean (SD); categorical: n (%). BMI = body mass index; CVD = cardiovascular disease.

Table 4. Baseline characteristics by HbA1c trajectory class: reduced dentition analysis (n = 572).

Variable	Overall	Low-stable	Moderate	High increasing
N	572	479	44	49
Age, years	61.7 (10.0)	62.9 (9.3)	54.9 (11.7)	55.9 (11.0)
Male	344 (60.1%)	290 (60.5%)	24 (54.5%)	30 (61.2%)
Ethnicity				
Chinese	372 (65.0%)	327 (68.3%)	25 (56.8%)	20 (40.8%)
Indian	93 (16.3%)	71 (14.8%)	7 (15.9%)	15 (30.6%)
Malay	72 (12.6%)	54 (11.3%)	7 (15.9%)	11 (22.4%)
Others	35 (6.1%)	27 (5.6%)	5 (11.4%)	3 (6.1%)
Diabetes duration, years	8.5 (7.7)	8.3 (7.5)	8.8 (9.4)	9.9 (8.4)
BMI, kg/m²	26.8 (4.8)	26.7 (4.8)	27.9 (4.7)	27.4 (5.0)
CVD history	150 (26.2%)	128 (26.7%)	11 (25.0%)	11 (22.4%)
Baseline teeth (mean)	23.0 (6.3)	22.9 (6.3)	23.0 (6.6)	24.0 (5.8)
Reduced dentition	180 (31.5%)	152 (31.7%)	13 (29.5%)	15 (30.6%)

Continuous variables: mean (SD); categorical: n (%). BMI = body mass index; CVD = cardiovascular disease.

Table 5. Trajectory-specific risk estimates [95% CI]: severe periodontitis (n = 1614).

Panel A: Risk estimates

Estimator	Low-stable	Moderate	High-increasing
Hard	0.734 [0.711, 0.757]	0.757 [0.679, 0.834]	0.782 [0.719, 0.845]
Soft	0.735 [0.712, 0.758]	0.748 [0.689, 0.807]	0.780 [0.727, 0.833]
TCCE (λ=0)	0.735 [0.712, 0.758]	0.747 [0.694, 0.800]	0.778 [0.728, 0.828]
TMLE–SL	0.740 [0.716, 0.763]	0.734 [0.654, 0.814]	0.753 [0.680, 0.826]

Panel B: Contrasts

Estimator	Contrast	RD [95% CI]	RR [95% CI]
Hard	High vs Low	0.048 [−0.022, 0.117]	1.065 [0.973, 1.166]
Soft	High vs Low	0.046 [−0.013, 0.104]	1.062 [0.985, 1.146]
TCCE (λ=0)	High vs Low	0.043 [−0.012, 0.098]	1.059 [0.985, 1.138]

TMLE–SL	High vs Low	0.014 [−0.067, 0.094]	1.018 [0.914, 1.134]
Hard	Mod vs Low	0.022 [−0.060, 0.104]	1.030 [0.923, 1.150]
Soft	Mod vs Low	0.014 [−0.050, 0.077]	1.018 [0.935, 1.110]
TCCE (λ=0)	Mod vs Low	0.013 [−0.044, 0.069]	1.017 [0.942, 1.098]
TMLE–SL	Mod vs Low	−0.006 [−0.091, 0.079]	0.992 [0.883, 1.115]

652

653 *Results pooled across M = 10 imputations using Rubin's rules. CI = compatibility interval.*

654 *λ = denominator shrinkage parameter.*

655

656 **Table 6. Trajectory-specific risk estimates [95% CI]: reduced dentition (n = 572).**

657 **Panel A: Risk estimates (adjusted for baseline tooth count)**

Estimator	**Low-stable**	**Moderate**	**High increasing**
Hard	0.317 [0.277, 0.358]	0.295 [0.158, 0.433]	0.306 [0.177, 0.435]
Soft	0.319 [0.278, 0.360]	0.316 [0.204, 0.427]	0.275 [0.171, 0.379]
TCCE (λ=0)	0.319 [0.278, 0.359]	0.316 [0.211, 0.420]	0.280 [0.182, 0.377]

TMLE–SL	0.309 [0.271, 0.347]	0.338 [0.264, 0.411]	0.364 [0.252, 0.476]

Panel B: Contrasts

Estimator	**Contrast**	**RD [95% CI]**	**RR [95% CI]**
Hard	High vs Low	−0.011 [−0.147, 0.124]	0.965 [0.611, 1.523]
Soft	High vs Low	−0.044 [−0.154, 0.065]	0.861 [0.574, 1.291]
TCCE (λ=0)	High vs Low	−0.039 [−0.142, 0.064]	0.877 [0.605, 1.271]
TMLE–SL	High vs Low	0.055 [−0.054, 0.164]	1.179 [0.831, 1.534]
Hard	Mod vs Low	−0.022 [−0.163, 0.119]	0.931 [0.563, 1.541]
Soft	Mod vs Low	−0.003 [−0.117, 0.111]	0.990 [0.682, 1.438]
TCCE (λ=0)	Mod vs Low	−0.003 [−0.110, 0.103]	0.989 [0.700, 1.399]
TMLE–SL	Mod vs Low	0.029 [−0.039, 0.097]	1.093 [0.887, 1.347]

Adjusted for baseline continuous tooth count in addition to the standard confounders.

Results pooled across M = 10 imputations using Rubin's rules.

Figure legends

Figure 1. Sample selection for the linked SDR-ODR analysis.

663 **Figure 2.** Overall mean RMSE for the four estimators across the three simulation scenarios.

664 **Figure 3.** Mean HbA1c trajectories for the three latent classes in the SDR-ODR cohort (n

665 = 1614)

Figures

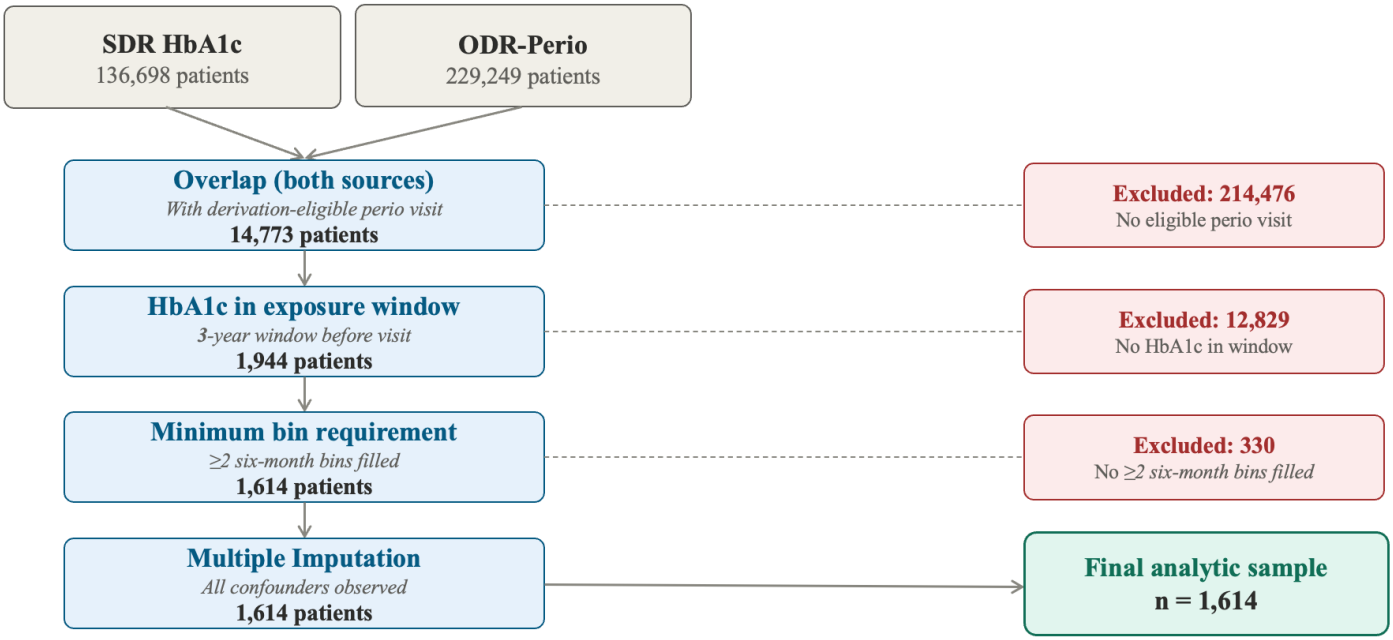


Figure 1

Sample selection for the linked SDR-ODR analysis.

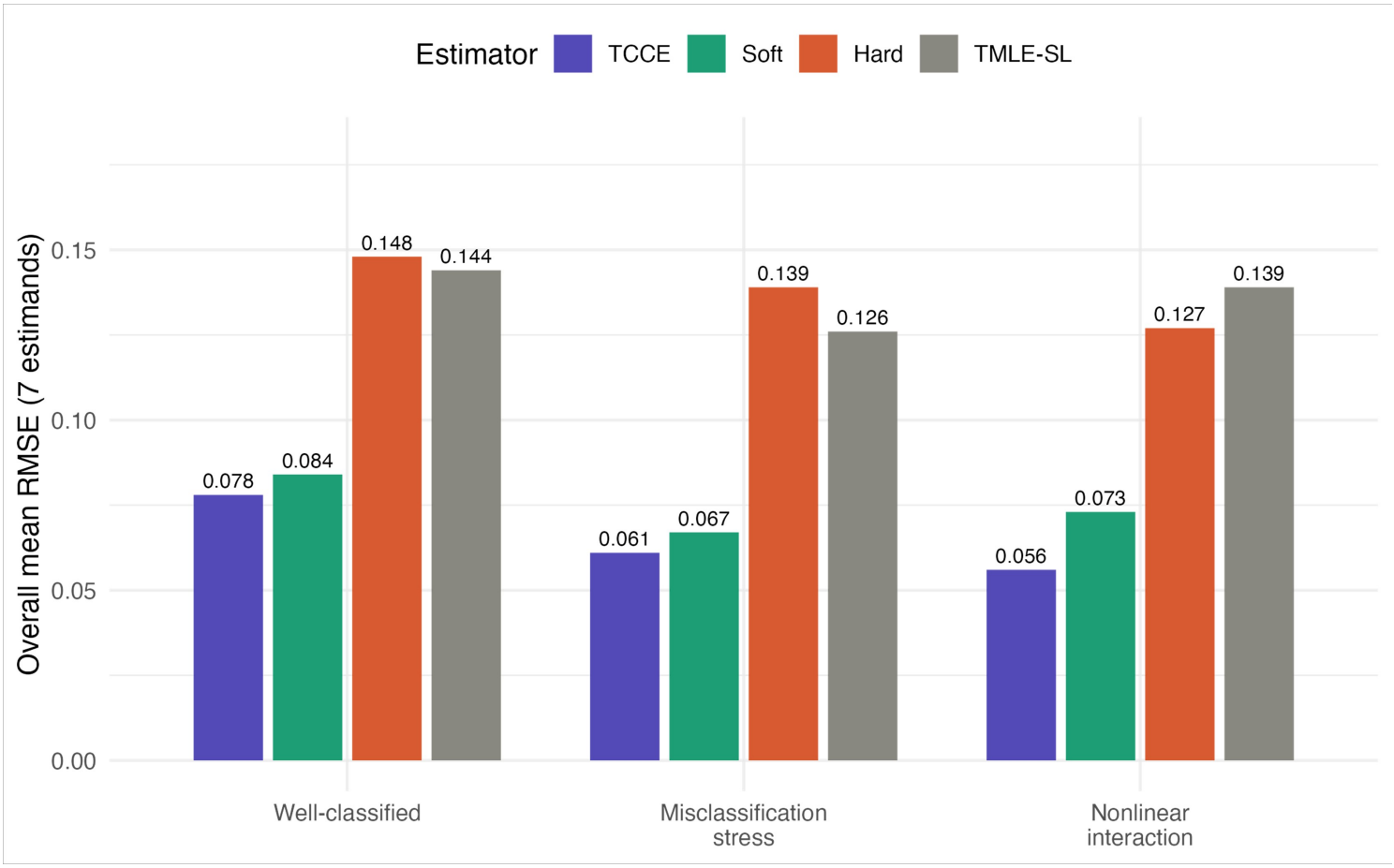


Figure 2

Overall mean RMSE for the four estimators across the three simulation scenarios.

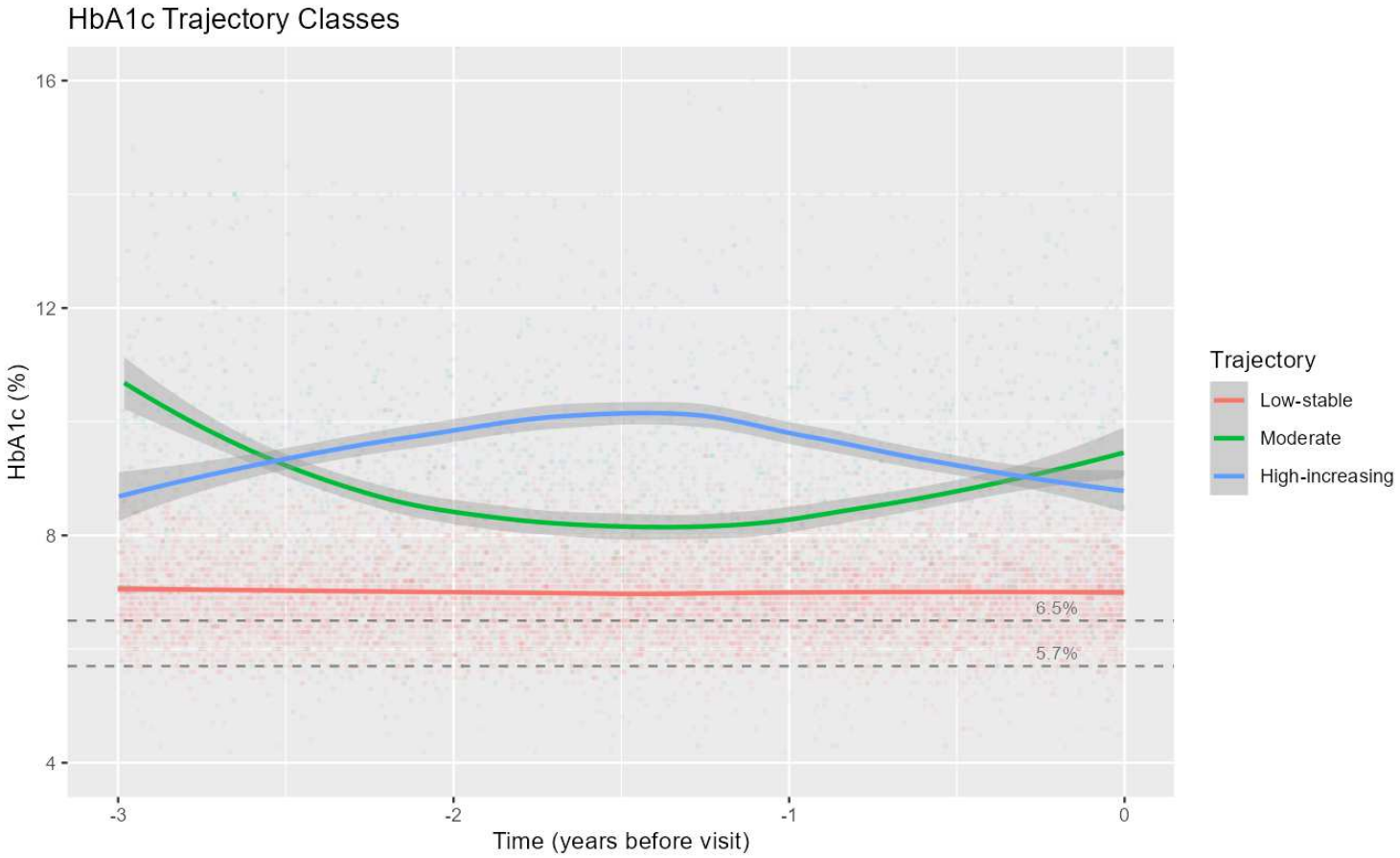


Figure 3

Mean HbA1c trajectories for the three latent classes in the SDR-ODR cohort (n = 1614)

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- TCCEsupplementarymaterial.docx