A Bayesian approach to common models of life course epidemiology

Justin Chumbley[1], Wenjia Xu[1], Cecilia Potente[1], Kathleen Mullan Harris[2], & Michael Shanahan[1]

[1] Jacobs Center for Productive Youth Development, University of Zurich

[2] Carolina population center, University of North Carolina

Author Note

Correspondence concerning this article should be addressed to Justin Chumbley, Jacobs Center for Productive Youth Development, Andreasstrasse 15, CH-8050 Zurich. E-mail: justin.chumbley@pm.me

Abstract

Background: Life course epidemiology studies people's health over long periods of time, treating repeated measures of their experiences (usually risk factors) as covariates or causes of subsequent morbidity and mortality. Three basic hypotheses often guide the analyst in modeling the repeated assessments of experience: accumulation (all measurement occasions have the same importance with respect to the predicted variable), critical-period (only one measurement occasion is important), and sensitive-period (more than one measurement occasion is of some meaningful importance). This last possibility is a vast composite hypothesis, referring to all possibilities that do not reflect accumulation or sensitive-period.

Methods: We propose two novel Bayesian quantities to test these three hypotheses, quantities that (a) contrast these three broad hypotheses against one another and (b) in the case of the sensitive-period hypothesis, identify which specific periods are relatively important. We test the approach via simulations, before presenting a real example relating obesity to a RNAseq measure of colorectal cancer disposition.

Results: The approach correctly identifies the life course hypothesis under which the data were simulated. Results from the empirical cohort study indicated with 97% probability that colorectal cancer disposition was more sensitive to recent obesity than obesity on any previous occasion. Historical obesity nevertheless still mattered: a sensitive model was more likely than any critical model.

Conclusions: The methods we present here help simplify the comparison of life course hypotheses prior to more formal analyses of causal effect.

Keywords: Developmental epidemiology, regression, life course models, ...

Word count: X

A Bayesian approach to common models of life course epidemiology

Key messages

- Life course epidemiological methods often test or compare points in parameter space by specifying multiple models and/or multivariate credible sets.
- We instead propose comparing realistic *regions* of parameter space, using a simple univariate reparameterization of a single "encompassing" model.
- We also describe the first general method for coherently decomposing the broad "sensitive hypothesis" into more specific, scientifically meaningful assertions.

Introduction

Life course epidemiology typically examines associations between repeated measures of "exposures" (usually risk factors) and subsequent, health-related variables over many decades of life. This approach has proven quite popular in the study of chronic forms of morbidity, which are rarely instantaneous results of relatively time-delimited exposures to risk factors, but rather often involve multiple exposures over many years and an appreciable latency period between exposure and outcome (Lynch & Smith, 2005). This body of research is usually guided by three mutually-exclusive life course models or hypotheses (Ben-Shlomo & Kuh, 2002): multiple exposures of roughly equal importance in predicting an outcome (the accumulation model); one of multiple exposures being decisively important (the critical period model); and several exposures predicting the outcome in non-trivial ways (the sensitive period model). Adjudicating among these models — as well as several others that are sometimes recognized (see Kuh & Shlomo (2004), Table 1) - is thus a fundamental task of life course epidemiology. In this paper, we present a Bayesian framework that builds on prior suggestions (Madathil, Joseph, Hardy, Rousseau, & Nicolau, 2018) and provides an effective, easily interpretable way to decide which of these models is most likely true given the data and, for sensitive period models, to order measurement occasions by how well they predict the outcome.

We assume that a subject's exposure history $\mathbf{x} = (x_1, ..., x_T)$ over T periods impacts their subsequent outcome Y without time-dependent confounding according to a generalized linear model with conditional expectation of the form $E(Y|\mathbf{x}) = g(\mathbf{x}\boldsymbol{\theta}')$. (Here g is some link function and linear covariates have been omitted without loss of generality.) The parameter $\boldsymbol{\theta} = (\theta_1, ..., \theta_T)$ captures the impact of exposure x at each time period $t = 1, ..., T$ on outcome Y, and we would like to contrast these effects to one another to see which measurement occasions(s), if any, matter more than others. Such contrasts represent tests of the accumulation and critical and sensitive period hypotheses. Madathil et al. (2018) propose a non-linear reparameterization $\boldsymbol{\theta} = \delta\mathbf{w}$ where $\delta \in \mathbb{R}$ and $\mathbf{w} \in \Delta^T$. Here $\Delta^T := \{(w_1, .., w_T) \in \mathbb{R}^T : w_t > 0, \sum_t w_t = 1\}$ is the T-part regular simplex, which has $T - 1$ degrees of freedom. This parameterization forces all non-zero weights to share the same sign, which is often a reasonable assumption in epidemiological studies of, for example, risk factors. Importantly, this approach gives readily interpretable parameters: w_t now reflects the relative effect of time period t while δ is “the total lifetime effect”. (Note that non-identity link functions g will generally also affect the interpretation of $\mathbf{w}$.)

Madathil et al. Madathil et al. (2018) assume a uniform prior on the weights $\mathbf{w}$. To compare the three broad hypotheses above, they ask whether their multivariate posterior 95% credible regions exclude or include (cover) the accumulation or critical hypotheses (points): these points are, respectively, a) *accumulation* $w_t = 1/T$ for all t, for example $\mathbf{w} = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ in the $T = 3$ period situation, and b) *critical period* $w_t = 1$ for some unique t, for example $\mathbf{w} = (0, 0, 1)$. Their strategy accommodates a wide variety of exposure and outcome variables, missing values, and measurement errors, in addition to other advantages of the Bayesian approach (e.g. Robert, 2007).

Yet limitations of continuous, multidimensional credible sets create acute challenges in bounded parameter spaces like Δ^T. Depending on the care taken in their construction and interpretation, multivariate credible sets may falsely exclude the critical model, which is found on the boundary of parameter space. They may also falsely include the accumulation model if practitioners erroneously conclude that marginal covering implies joint covering. In fact, a multivariate set S may exclude a point p even while all of P's lower dimensional projections include all of p's projections: see supplementary Figure 1 for some intuition. A univariate alternative would be both intuitive and side-step such complications with the treatment of multivariate credible sets. More importantly, even when the accumulation and critical point hypotheses are appropriately excluded by a posterior credible set, no conventional credible set can answer the question: "which specific periods are more sensitive than others?".

Building on Madathil et al. (2018)'s "continuous model expansion," we address these issues by constructing credible sets based on two *transformations* of their posterior distribution $p(\mathbf{w}|y)$. The first transformation seeks a simple univariate way to compare the three hypotheses, the second to decompose and interpret the composite sensitive period hypothesis. The first transformation calculates the greatest difference between any two period's weights, i.e. the range of component parameters $\mathbf{w} = (w_1, ..., w_T)$. This range characterizes the three hypotheses: it equals zero and one for the accumulation and critical hypotheses, respectively, and is strictly between zero and one for any sensitive model. The range therefore contains all the information required to choose between hypotheses and offers a practical alternative to conventional model selection which typically rests on the Bayes factor. (Alternatively, this transformation may be viewed as a quantitative index of how similar the weights are across time periods.) The second transformation ranks measurement occasions by relative importance of magnitude, assessing which periods are more "sensitive" (have larger weight) than others. This transformation permits a practitioner to assign a numeric probability to a conclusion such as "the first two periods of life matter more than any subsequent period."

Rationale

Our first transformation is $\phi : \Delta^T \to \Delta^2$ from the T-part simplex to the 2-part simplex or unit interval $[0, 1]$, defined by $\phi : \mathbf{w} \mapsto (\vee\mathbf{w} - \wedge\mathbf{w})$. Here $\vee$ is the max operation that extracts the magnitude of the largest component of $\mathbf{w}$, dually for the min operation $\wedge$. The function ϕ therefore gives the range of components of $\mathbf{w}$: the Euclidean distance between the maximum and minimum values of the components of $\mathbf{w}$. Note that $\phi((\frac{1}{3}, \frac{1}{3}, \frac{1}{3}) = \frac{1}{3} - \frac{1}{3} = 0$ for the accumulation hypothesis and $\phi((1, 0, 0)) = \phi((0, 1, 0)) = \phi((0, 0, 1)) = 1 - 0 = 1$ for any critical hypothesis. Note also, for example, that $\phi((0.2, 0.7, 0.1)) = 0.6 = 0.7 - 0.1$. This particular ϕ value of 0.6 represents a sensitive hypothesis because it is neither 0 nor 1 (pure accumulation or critical hypotheses). Other solutions to $\phi(\mathbf{w}) = 0.6$ represent sensitive hypotheses which are equivalent to $(0.2, 0.7, 0.1)$ in as much as they have the same range, including for example $\phi((0.8, 0.2, 0)) = 0.6$.

More generally, any choice of two distinct thresholds $a, b \in [0, 1]$ partitions the $\mathit{codomain}(\phi) = [0, 1]$ into three sets practically equivalent to the $\mathit{accumulation} = [0, a]$, $\mathit{sensitive} = (a, b)$ and $\mathit{critical} = [b, 1]$ hypotheses, respectively. (In principle this can be further generalized to include any finite number of break points $a_i \in [0, 1]$, yielding intermediary classes.) The strictest definition of these three hypotheses arises in the limit that a tends to 0 and b tends to 1. Note that $p(\mathit{sensitive\ or\ accumulation\ or\ critical}|y) = 1$ so this approach effectively discretizes $p(\phi|y)$ or equivalently. (Note that such a partition of the codomain exactly corresponds to a partition of $\mathit{domain}(\phi) = \Delta^T = M_0 \cup M_1 \cup M_2 := \{\mathbf{w} : 0 \leq \phi \leq a\} \cup \{\mathbf{w} : a < \phi < b\} \cup \{\mathbf{w} : b \leq \phi \leq 1\}$. Just as the posterior density of $p(\phi|y)$ at the value 0.6 reflects the total density over all these solutions $\{\mathbf{w} : \phi(\mathbf{w}) = 0.6\}$, our tripartition of $\Delta^T = M_0 \cup M_1 \cup M_2$ effectively discretizes continuous posterior density $p(\mathbf{w}|y)$ satisfying $1 = \int_{\Delta^T} p(\mathbf{w}|y)$ into discrete density $P(M_i|y)$ satisfying $1 = \sum_{i=1}^{3} P(M_i|y)$.) $p(\mathbf{w}|y)$. If $p(\mathit{sensitive}|y) \geq 0.95\%$, for example, we may conclude that the sensitive model is credible in practice.

Our second transformation then seeks to whittle down this broad sensitive period hypothesis into a more specific about the relative importance of time periods. This transformation is $f : \Delta^T \rightarrow \mathcal{S}_T$ from the simplex to the set of all full rankings (i.e. permutations of the labels {"w_1","w_2",. . . ,"w_T"} or equivalently of $\{1, 2, ..., T\}$). In particular, f assigns each vector $\mathbf{w}$ the full ranking of its components. For example, f maps the point $\mathbf{w} = (0.2, 0.7, 0.1)$ to the full ranking $w_3 < w_1 < w_2$: we will henceforth use the terser notation $3|1|2$ for such full rankings. Other solutions to $f(\mathbf{w}) = 3|1|2$, such as $\mathbf{w} = (0.3, 0.5, 0.2)$, represent sensitive hypotheses which are equivalent to $(0.2, 0.7, 0.1)$ in as much as the relative importance of periods is exactly the same. In this way we can again discretize continuous posterior density $p(\mathbf{w}|y)$, this time into the discrete density $P(\pi|y)$ over $T!$ full rankings π such as $3|1|2$ satisfying $1 = \sum_{\pi \in \mathcal{S}_T} P(\pi|y)$. This enables us to answer directly whether the most probable ranking of periods by importance is say $\pi = 3|1|2$, or whether say $p(\pi|y) \geq 0.95\%$. In this way we gain insight into our multivariate posterior $p(\mathbf{w}|y)$ without the inconvenience of complicated continuous multivariate credible sets.

The function f also allows one to define more general, *partial rankings* such as $3, 1|2$ and calculate their posterior probability. Such partial rankings denote collections of full rankings consistent with a weaker conclusion than a single full ranking on its own. For example, if $\pi = 1, 3|2$ and $p(\pi|y) \geq 0.95\%$ then we can say with 95% posterior credibility that period 2 is more important than the other two periods, even though we can say nothing about the relative importance of these latter. To elaborate, the partial ranking $1, 3|2$ represents points $\mathbf{w}$ that can be ranked *either* as $w_3 < w_1 < w_2$ or as $w_1 < w_3 < w_2$: it therefore encodes points $\mathbf{w}$ for which w_2 is unambiguously the most important period. However this ordering is *partial* because either $w_1 < w_3$ or $w_3 < w_1$. The set of all symbols like $1, 3|2$, which includes T integers separated by either a bar "|" or a comma "," covers a large set of scientific statements that are both readily interpretable and can be assigned posterior probability. In this framework, the $\beta\%$ finest credible rank, which we denote $\mathcal{C}_\beta$, is the smallest such set with $\beta\%$ posterior probability.

Goals of simulation

We ask whether the univariate credible interval for $\delta|y$ appropriately excluded zero, i.e. correctly inferred whether *any* time period is relevant for the outcome. If the answer to this is positive, it makes sense to broadly examine the three competing hypotheses (critical, accumulation and sensitive) via the posterior distribution $\phi|y$. If $\phi|y$ additionally supports a sensitive hypothesis, one should examine $f|y$ and the probable ordering of weights. The objectives of the simulation study were therefore: (a) to assess whether $\delta|y$ appropriately excluded zero and, if so, (b) to assess whether the range $\phi|y$ successfully discriminates between the accumulation, critical, and sensitive hypotheses; and in the case of the last possibility (c) to assess whether rank $f|y$ succeeds in identifying the correct ranking of time periods by their sensitivity.

Knowing the simulated ground truth $\mathbf{w}^*$ and its corresponding true full ranking $f(\mathbf{w}^*)$ our principle questions concern its relation to the inferred *finest β% credible ranking* which we denote $\mathcal{C}_\beta$.

1) Is $\mathcal{C}_\beta$ consistent with the true full ranking, i.e. $f(\mathbf{w}^*) \in \mathcal{C}_\beta$? We say that $\mathcal{C}_\beta$ is inconsistent if, for example, it asserts $w_2 < w_4$ while the underlying truth is $w_4 < w_2$. Otherwise it is consistent.

2) How much "information" does $\mathcal{C}_\beta$ retain? Here we use $q = r/r^*$ with values between 0 to 1 to measure the quality of $\mathcal{C}_\beta$, where r is the number of distinctions (inequalities or bars "|") in $\mathcal{C}_\beta$ and r^* the true number in $f(\mathbf{w}^*)$. Larger q therefore means a more informative inference.

The first question expresses the minimal requirement that $\mathcal{C}_\beta$ does not contradict the truth. The second question is motivated by the desire that $\mathcal{C}_\beta$ be as informative as possible, ideally faithfully retaining *all* distinctions made in the true ranking $f(\mathbf{w}^*)$.

Simulation parameters

Our simulation fully reproduced and extended that of Madathil et al. (2018). Namely, we simulated a three-period life course study assuming no measurement error in the variables. In particular, for participant i we sampled three Gaussian exposure variables $\mathbf{x}_i = (x_1, x_2, x_3)$ with a correlation of 0.7 and 0.49 between adjacent and non-adjacent measures, respectively. Datasets were simulated for all combinations of the four life course hypotheses (the two distinct sensitive hypotheses defined below, in addition to the accumulation and critical hypothesis) and three sample sizes ($n = 700, 1500, 3000$). The ground truth weight values of the simulation, denoted with an asterix "$*$," were: (i) pure accumulation hypothesis $\mathbf{w}^*_i = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$; (ii) monotonic sensitive period model with weights $\mathbf{w}^*_{ii} = \frac{1}{1+2+3}(1, 2, 3)$; (iii) first and second measurement occasions as a sensitive period $\mathbf{w}^*_{iii} = (0.75, 0.2, 0.05)$; (iv) third measurement occasion as a critical period $\mathbf{w}^*_{iv} = (0, 0, 1)$.

We extended Madathil et al. (2018)'s 3 period simulation to 5 and 7 periods as follows: The accumulation model (i) above was generally $\mathbf{w}^*_i = (\frac{1}{T}, ..., \frac{1}{T})$; (ii) was generalized in the obvious way to $\mathbf{w}^*_{ii} = \frac{1}{\sum_{i=1}^{T} t}(1, 2, ..., T)$ and (iii) and (iv) were padded with zeros, being for example $\mathbf{w}^*_{iii} = (0.75, 0.2, 0.05, 0, 0, 0, 0)$ and $\mathbf{w}^*_{iv} = (0, 0, 0, 0, 0, 0, 1)$ respectively in 7 dimensions.

We independently varied the lifetime effect δ^* between 0,1 and 2. This simulates situations where a unit change in the total exposure (or weighted average exposure) over all time points increases the outcome by 0, 1, or 2 units. These fixed underlying settings (estimands) are again denoted "$*$," to distinguish them from the posterior inferred counterparts. Given $\delta^*, \mathbf{w}^*$, we then generated $y_i = \delta^* \sum_{j=1}^{T} x_{ij} w^*_j + \epsilon_i$ with independent $\epsilon_i \sim N(0, 1)$, for $i = 1, ..., n$.

Prior, likelihood and posterior

In accordance with the data generating model above, we used Bayesian linear regression for inference. We followed Madathil et al. (2018) in their choice of a uniform prior over all Δ^T, namely a non-informative Dirichlet prior for weights $p(\mathbf{w}) = Dirichlet(\mathbf{w}|\mathbf{1})$, where $\mathbf{1}$ is a vector of T ones. In cases where there is plausible justification for bias towards the accumulation or critical models, the hyperparameter can be generalized to $\alpha\mathbf{1}$, with $\alpha > 0$. Then it is well-known from the properties of the Dirichlet that choosing $\alpha < 1$ implies a bias toward the critical hypothesis and $\alpha > 1$ a bias towards accumulation. We revisit these options below. We also chose a weakly informative Cauchy prior on the lifetime effect $p(\delta) = Cauchy(\delta|0, 2.5)$.

We then used the No-U-Turn MCMC sampler as implemented in STAN (Carpenter et al., 2017) to acquire 50k marginal posterior samples of $\delta|y$ and $\mathbf{w}|y$. Having performed standard convergence tests, we examined $\delta|y$, and derived $\phi|y$ and $f|y$ by applying ϕ, f to each point in our posterior sample $\mathbf{w}|y$.

(FIGURE 1 HERE)

Results of simulation

Posterior lifetime effect δ and range ϕ

In every simulation the univariate credible interval for $\delta|y$ appropriately excluded zero. Thus $\delta|y$ is a faithful omnibus measure of lifetime effect. In simulations with evidence of non-zero $\delta|y$, we proceeded to examine the posterior range $\phi|y$. The violin plots in Figure 1 depict this posterior distribution of the range $\phi|y$ under each setting of $\mathbf{w}^*$ the ground truth, the total life-time effect δ^*, the number of periods T, and the sample size n. The x-axis of each plot refers to different life course epidemiology models. i refers to simulations where the accumulation model is true, ii, iii indicate the sensitive models and iv indicates the critical model.

Note that the posterior distribution $\phi|y$ approaches the truth $\phi(\mathbf{w}^*)$ with increasing δ^* and n. To convert these plots into a more formal comparison between the *accumulation*, *critical-period* and *sensitive* hypotheses, we choose $a = 0.15$ and $b = 1 - a$ to define regions practically equivalent to each of these three cases, as discussed in section "Rationale". Table 1 reports the resulting confusion matrix which relates the inferred hypotheses (column variable) to the underlying truth (row variable). We made a conclusive choice between these cases whenever one of them had probability greater than 0.9 posterior probability, otherwise our inference was considered inconclusive/unknown (u). The table illustrates that we could faithfully recover the ground truth, albeit at the price of occasionally acknowledging that a conclusion is not possible given the data. There were no incorrect inferences, such as would arise if we concluded that the accumulation was true when in fact data was generated under the sensitive model, etc. We see from this table that around 90% (in this case exactly $(9)/(18 \times 4)$ of inferences were conclusive and correct. The remaining inferences were inconclusive. For example, of the 18 simulations where either the accumulation model was true (Table 1, row 1) inference was correct in 14 cases and inconclusive otherwise.

(TABLE 1 HERE)

When the sensitive hypothesis was credible, i.e. the 18+17 cases in column s of Table 1, it makes sense to ask which periods are more or less sensitive. For these cases we therefore calculated the finest partial ranking of parameters, as discussed next.

(FIGURE 2 HERE)

(TABLE 2 HERE)

Posterior finest credible rank $\mathcal{C}_\beta$

Figure 2 illustrates our recursive scheme for whittling down the sensitive hypothesis. In particular, it gives the "cumulative density function" of a special nested increasing set of subsets of Δ^T which leads to $\mathcal{C}_{90\%}$, the finest 90% credible ranking (see Figure 2). The choice of which distinction is weakest is determined by a recursive maximization scheme.

The candidate partial ranking at each step in the sequence from left to right is the most credible (maximum probability) coarsening of the preceding candidate. In practice, the posterior probability of each candidate ranking was estimated as the fraction of posterior samples satisfying the relevant inequalities. We selected $\mathcal{C}_{90\%}$ as the first candidate that exceeds the desired credibility threshold of 0.9, depicted in Figure 2 as a black horizontal line. The column "fcr" of Table 2 gives some examples of this inferred finest credible ranking $\mathcal{C}_{90\%}$ across different simulations. Take row 1 for example. The finest credible ranking was $1, 2|3$ which means that measurement 3 was credibly more weighty than both the others, which were indistinguishable. The simulated ground truth in row 1 was in fact a monontonic increase in the weight of exposure with measurement occasions: the true model was model *ii* and the true ranking ("truth") was $1|2|3$. Evidently, while the the fcr $\mathcal{C}_{90\%}$ of $1, 2|3$ does not contain complete information about the underlying truth (it is uncertain of any distinction between time point 1 and 2) neither does it contradict or violate this truth. The latter would arise if, for example, the fcr were say $1|3|2$ or $2, 3|1$, etc. Then there would be at least on pair of measurement occasions whose asserted importance or ordering in the fcr contradicts (is opposite to) the underlying truth. Because the fcr in row 1 makes only 1 distinction of a total of 2, thus the value 1/2 in the column "q".

This fcr was systematically compared with the "truth" column over all simulations to answer the two specific questions posed in section "Goals of Simulation:"

1) We found that our inferred partial rank $\mathcal{C}_\beta$ rarely violated the ground truth. Such a

violation occurred in 0 percent of the simulations.

2) On average over all simulations 0.71 % of the distinctions were preserved. Table 3 shows that q, the proportion of distinctions preserved in $\mathcal{C}_\beta$, increased with the simulated sample size. On average over simulations with the lowest sample of 700, 52% of distinctions were preserved. This raised to 72% and 89% with higher sample sizes. This shows how increased posterior precision in $\mathbf{w}|y$ translates to increased precision of the inferred rank.

(TABLE 3 HERE)

Empirical example

Data and regression model

The data come from the National Longitudinal Study of Adolescent to Adult Health (Add Health), a representative study of US adolescents in grades 7-12 in 1994-95 who have been followed into adulthood over five waves of data collection (Harris, 2013). Specifically, the present study combines body weight data from a) Birth records (0 years) b) Waves I and II (12-20 years) c) Wave III (18-27 years) d) Wave IV (25-33 years) e) Wave V Sample 1 (33-42 years).

RNASeq abundance data from peripheral blood samples was collected at Wave V (currently, n=1132 samples collected in the 2016-2017 window have been fully processed). Wave 1 and wave 2 were very close in time (1995 and 1996 respectively) and so have been pooled to miminize missing data: in practice, all but 12 subjects data come from wave I.

The goal is to predict a pre-symptomatic mRNA colorectal cancer signature from five measurements of pronounced body weight, taken on 5 separate measurement occasions:

pronounced birth weight (defined as being birth weight higher than 8.8 pounds or lower than 5.5 pounds) and then obesity (bmi ≥ 30) on waves I-V. We would like to know whether one measurement occasion is *critical* for predicting variation in this cancer signature, if all occasions matter equally (*accumulation*), or whether multiple occasions are sensitive albeit to different degrees (*sensitive*). In the latter case we would like to identify the most confident statement about the relative sensitivities possible.

Our outcome variable was a scalar mRNA colorectal cancer signature constructed as the normalized, weighted mean of 127 up-regulated genes (given positive weight) and 2 down-regulated genes (given negative weight). These genes are collectively implicated in colorectal cancer biology (Guinney et al., 2015). Normalization was performed using a reference-gene normalization procedure (Eisenberg & Levanon, 2013) and converts raw counts onto a log scale. We had both mRNA and phenotype data for a total of 893 participants. For participant i we denote the "pronounced body weight" dummies on the five measurement occasions $\mathbf{x}_i = (x_1, x_2, x_3, x_4, x_5)'$, and their real-valued mRNA score $\bar{y}_i$. Our models controlled for biological sex, race/ethnicity, age at time of the survey at wave 1, preterm birth status, current smoking, binge drinking, region, and sample-specific quality control measures for mRNA. For notational simplicity, these control variables are denoted by the vector $\mathbf{c}$ in the equations below.

We first used a Bayesian MCMC approach to estimate the reparameterized regression model of Madathil et al. (2018) as $\bar{y}_i = \alpha + \delta \sum_{i=1}^{5} w_i x_i + \mathbf{c}'\gamma + \epsilon_i$. Here α is an intercept, δ and $\mathbf{w} = (w_1, ..., w_5)$ are as defined throughout this work, and γ is the regression parameter vector for our control variables. Assuming a uniform prior distribution on $\mathbf{w}$ and weakly informative normal prior distributions on δ and γ, we used MCMC to collect 50k samples from their posterior distributions using the R package "rstan", as implemented by Guo et al. (2018). In particular, we pooled post-burn in samples from 5 independent chains of 10k having performed standard diagnostics to ensure convergence (we run 5 independent parallel

Hamiltonian Monte Carlo chains, each with 20k iterations and convergence was confirmed by checking trace plots and Rhat values).

Results

The central 90% and 95% posterior intervals for the lifetime effect δ were respectively [0.09, 0.17] and [0.08, 0.18]. These exclude zero, so we proceeded to evaluate the suitability of accumulation, critical and sensitive models by examining the posterior range $\phi|y$. Recall that ϕ is defined to be the difference between the largest and smallest components of $\mathbf{w}$. Following our **Rationale** above, we chose two thresholds $a = 0.15, b = 0.85$ to partition the possible values of ϕ into three non-overlapping intervals, corresponding to our three life course "models" of interest. We found the posterior probabilities for the *accumulation*, *sensitive* and *critical* models were 0.00046, 0.99208 and 0.00746 respectively. The sensitive model was most credible so we asked which periods were more sensitive than which by examining the probable ranking of weights. Table 4 gives the highest probability partial rankings at each level of granularity. The most probable *full* ranking was 3|4|2|1|5 but unfortunately this highly informative statement carried only 15% credibility. In fact the finest credible ranking, with 97% credibility, was $3, 4, 2, 1|5$ which can equally be written $1, 2, 3, 4|5$ in our notation (because the order of numbers *between* any two adjacent bars is arbitrary in our notation). Thus the final, fifth measurement occasion matters more than any other (birth or waves I-IV) in predicting outcome. We have already concluded that (some of) these early periods probably matter: the posterior probability that they all have weight zero (i.e. the probability of a *critical* pattern of weights) is only 0.00746, as given above. Yet our inferential uncertainty in this data set is too high to warrant any further claims about the relative importance of these earlier measurement occasions. While this may appear unsatisfactory, it is simply an honest consequence of our Bayesian strategy of carefully handling and transparently integrating inferential uncertainty.

Discussion

We have presented a novel way to assess three commonly used models of life course epidemiology: the critical, accumulation and sensitive models. Our approach rests on two reparameterizations of the model described by Madathil et al. Madathil et al. (2018). Faced with the challenge of interpreting multivariate weights and credible sets, our methods exploit novel and concise univariate summaries. They therefore aid interpretability and reportability, offering principled *qualitative* inferences both about which model is "best" (accumulation, critical or sensitive) and in the latter case, which periods are relatively more sensitive. These generalizations of (Madathil et al., 2018) are achieved by relaxing the dependence on both continuous multivariate confidence sets and the chosen metric on parameter space (previously chosen to be the Euclidean metric as opposed to say Hilberts projection metric or the Aitchison metric on the simplex). Our decomposition of the sensitive model does not require any metric, being based instead on order relations. Crucially however, our approach remains true to their spirit of "continuous model expansion:" we have not needed to explicitly specify multiple models or to use the Bayes Factor for formal comparison. Rather, we showed that one can simply summarize the posterior parameters of a single encompassing regression model. Our approach may be particularly attractive in situations of high prior uncertainty about the true relative sensitivities. In such situations, there is no compelling reason to artificially privilege the critical and accumulation points in parameter space with prior probability mass, as required by model comparison methods based on the Bayes factor. Equally, assuming the sensitive model were true, there may be no compelling reason to to privilege one pattern of temporal sensitivity to exposure over another. In such cases it makes sense to place a uniform distribution over parameter space and therefore over different patterns (full rankings) of sensitivity. For a detailed discussion of alternative priors on ϕ and f, see the supplementary material. Our approach may be less appropriate for modelling protracted exposures with closely spaced measurements VACEK (1997), but see Madathil et al. (2018).

Our two proposed methods express posterior uncertainty differently. We considered our *model adjudication* inconclusive if no single model (sensitive, critical and accumulation) attained some desired posterior probability, say 90%. (Alternatively, we could have used the width and location of the range ϕ to describe uncertainty via a univariate credible interval.) In contrast, we considered our *sensitive model decomposition* completely inconclusive only when no relative order between any two measurement occasions was credible at the desired posterior probability of 90%. But the 90% credible conclusions of this method could generally vary on a spectrum from *partial* to *full*, reflecting an increasing proportion of bars "|" to commas "," in our chosen notation for partial rankings. Our simulations illustrated that uncertainty of both our methods reduced with sample size and increased with effect size. Importantly, neither method was over-confident: model adjudication never confidently selected the wrong model, and sensitive model decomposition rarely confused the relative importance of measurement occasions. We therefore used our methods to analyse a real data example from the Add health cohort. In particular, we asked how current mRNA cancer disposition reflected the history of body weight. We found in this case that the critical and accumulation models could not offer a compelling explanation for the data. Instead a sensitive pattern emerged in which mRNA was most sensitive to the most recent measurement occasion, but historical body weight still made some difference.

Future work should extend our methodological approach to time-dependent parameters of more explicitly causal models, e.g. (Robins, Hernan, & Brumback, 2000; VanderWeele, Hernán, Tchetgen Tchetgen, & Robins, 2016).

Acknowledgements

We would like to thank Margaret Bellamy.

Supplementary material

Prior on ϕ and f. Our first methodological proposal offers a new way to choose between accumulation, critical and sensitive models. In contrast to our proposed strategy, and that of Madathil et al. (2018), a more conventional approach to model selection using Bayes factors would require explicitly specifying point mass priors on the critical and accumulation points, together with a continuous prior over the sensitive hypothesis. As explained below, this conventional approach mandates a dramatic increase in the prior plausibility of the accumulation and critical cases *which are otherwise quite implausible on mathematical grounds alone.* It is partly to avoid artificially distinguishing the two point hypotheses in this way, and necessarily boosting their prior plausibility, that we follow Madathil et al. (2018). Like those authors we choose a uniform prior over all Δ^T, but we proceed by formally defining "near accumulation" or "near critical" models (not points) via two thresholds on the domain of our derived parameter ϕ, the *range* of weights. In either case, clearly prior uniformity over $\mathbf{w}$, implemented say by setting the Dirichlet hyperparameter to $\alpha = 1$, actually favors the sensitive model: it does *not* imply prior uniformity over the critical, accumulation and sensitive hypotheses. This simply reflects the mathematical fact that the accumulation and critical points occupy less of Δ^T, and therefore are assigned proportionally less prior probability. In fact, *interpreting the critical or accumulation hypotheses as points would imply they have zero prior and posterior probability under any continuous distribution*: technically speaking, only volumes and not individual points have non-zero probability in this setting. Conversely, the sensitive hypothesis naturally comprises the entire volume of parameter space Δ^T and hence has 100% prior and posterior probability. Of course, a uniform prior on the weights $\mathbf{w}$ might be inappropriate in applications with substantial prior scientific knowledge. Our approach can accommodate such applications where previous data or theory indeed support a critical or accumulation model *a priori*: by calibrating the Dirichlet hyperparameter the scientist may introduce prior bias of the desired strength towards either the accumulation model ($\alpha > 1$) or the critical model ($\alpha < 1$).

Our second methodological proposal was to decompose the sensitive model which yields the so-called "finest credible ranking" of measurement occasions by their impact on the outcome. Note that, unlike the case of model adjudication discussed in the previous paragraph, each constituent full ranking of the sensitive hypothesis, say $3|1|2$ or $3|2|1$, *occupies equal volume of* Δ^T. Unlike with model adjudication via ϕ, uniformity of $p(\mathbf{w})$ is therefore indeed preserved by the rank transformation f. In particular, each full ranking is assigned $1/T!$ prior probability. A general partial ranking is assigned prior probability $k/T!$, where k is the number of underlying full rankings that comprise the partial ranking. For example, partial ranking $2|1, 3$ has $k = 2$ and prior equaling $2/3!$ because $2|1, 3 := 2|1|3$ *or* $2|3|1$. In applications where there are clear competing scientific theories about the precise sensitivity profile *a priori*, it may be possible to use confirmatory methods suggested in the literature on so-called "informative hypotheses" (Gu, Mulder, Deković, & Hoijtink, 2014; Klugkist, Kato, & Hoijtink, 2005; Mulder & Olsson-Collentine, 2019; Mulder, Hoijtink, & Klugkist, 2010).

The ease with which we can flexibly pose and answer such general questions is a key motivation for taking the Bayesian path. By comparison, the frequentist linear modelling framework (Rosenthal, Rosnow, & others, 1985; Rosenthal, Rosnow, & Rubin, 2000) is limited to rejecting sensitivity hypotheses which can be cast as linear contrasts $\mathbf{w}^T\mathbf{c}$ for some *fully pre-specified* contrast vector $\mathbf{c}$. This traditional workhorse can therefore reject, for example, a fully specified linear or exponential sensitivity hypothesis. (These include: sensitivity linearly decays as $w_{t+1} = w_t + \alpha$ with fixed, hypothetical $\alpha < 0$, or exponentially as $w_t = \alpha^t$ with $0 < \alpha < 1$. In the former, $w_{t+1} - w_t = \alpha$ and in the latter $w_{t+1}/w_t = \alpha$ for all $t < T$.) Yet the science rarely justifies committing to a linear or exponential trend, let alone a strong hypothetical specification of α. Our proposed approach is not limited to rejecting null hypotheses. It can positively accept the weaker claim that sensitivity decreases in some monotone fashion $w_1 > w_2 > ... > w_T$. This latter includes the linear and exponential special cases but is less restrictive (and therefore more credible) because it doesn't insist on a particular functional form for the change in sensitivity over time. More generally still, it can identify arbitrary patterns of sensitivity $w_2 < \{w_3, w_1\} < w_4$ or $\{w_3, w_1\} < \{w_1, w_4\}$, etc. This lead us to develop the notion of the finest credible ranking $\mathcal{C}$ which may be derived automatically from the data.

REFERENCES

Ben-Shlomo, Y., & Kuh, D. (2002). A life course approach to chronic disease epidemiology: Conceptual models, empirical challenges and interdisciplinary perspectives. Oxford University Press.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., … Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, *76*(1).

Eisenberg, E., & Levanon, E. Y. (2013). Human housekeeping genes, revisited. *TRENDS in Genetics*, *29*(10), 569–574.

Gu, X., Mulder, J., Deković, M., & Hoijtink, H. (2014). Bayesian evaluation of inequality constrained hypotheses. *Psychological Methods*, *19*(4), 511.

Guinney, J., Dienstmann, R., Wang, X., De Reyniès, A., Schlicker, A., Soneson, C., … others. (2015). The consensus molecular subtypes of colorectal cancer. *Nature Medicine*, *21*(11), 1350.

Guo, J., Lee, D., Sakrejda, K., Gabry, J., Goodrich, B., De Guzman, J., … Fletcher, J.

(2018). RStan: The r interface to stan. r package version 2.17. 3.

Harris, K. M. (2013). The add health study: Design and accomplishments. *Chapel Hill: Carolina Population Center, University of North Carolina at Chapel Hill*.

Klugkist, I., Kato, B., & Hoijtink, H. (2005). Bayesian model selection using encompassing priors. *Statistica Neerlandica*, *59*(1), 57–69.

Kuh, D., & Shlomo, Y. B. (2004). *A life course approach to chronic disease epidemiology*. Oxford University Press.

Lynch, J., & Smith, G. D. (2005). A life course approach to chronic disease epidemiology. *Annu. Rev. Public Health*, *26*, 1–35.

Madathil, S., Joseph, L., Hardy, R., Rousseau, M.-C., & Nicolau, B. (2018). A bayesian approach to investigate life course hypotheses involving continuous exposures. *International Journal of Epidemiology*, *47*(5), 1623–1635.

Mulder, J., & Olsson-Collentine, A. (2019). Simple bayesian testing of scientific expectations in linear regression models. *Behavior Research Methods*, 1–14.

Mulder, J., Hoijtink, H., & Klugkist, I. (2010). Equality and inequality constrained multivariate linear models: Objective model selection using constrained posterior priors. *Journal of Statistical Planning and Inference*, *140*(4), 887–906.

Robert, C. (2007). *The bayesian choice: From decision-theoretic foundations to computational implementation*. Springer Science & Business Media.

Robins, J. M., Hernan, M. A., & Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. LWW.

Rosenthal, R., Rosnow, R. L., & others. (1985). *Contrast analysis: Focused* *comparisons in the analysis of variance*. CUP Archive.

Rosenthal, R., Rosnow, R. L., & Rubin, D. B. (2000). *Contrasts and effect sizes in behavioral research: A correlational approach*. Cambridge University Press.

VACEK, P. M. (1997). Assessing the effect of intensity when exposure varies over time. *Statistics in Medicine*, *16*(5), 505–513.

VanderWeele, T. J., Hernán, M. A., Tchetgen Tchetgen, E. J., & Robins, J. M. (2016). Re: Causality and causal inference in epidemiology: The need for a pluralistic approach. *International Journal of Epidemiology*, *45*(6), 2199–2200.

Ben-Shlomo, Y., & Kuh, D. (2002). A life course approach to chronic disease epidemiology: Conceptual models, empirical challenges and interdisciplinary perspectives. Oxford University Press.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., … Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, *76*(1).

Eisenberg, E., & Levanon, E. Y. (2013). Human housekeeping genes, revisited. *TRENDS in Genetics*, *29*(10), 569–574.

Gu, X., Mulder, J., Deković, M., & Hoijtink, H. (2014). Bayesian evaluation of inequality constrained hypotheses. *Psychological Methods*, *19*(4), 511.

Guinney, J., Dienstmann, R., Wang, X., De Reyniès, A., Schlicker, A., Soneson, C., … others. (2015). The consensus molecular subtypes of colorectal cancer. *Nature Medicine*, *21*(11), 1350.

Guo, J., Lee, D., Sakrejda, K., Gabry, J., Goodrich, B., De Guzman, J., … Fletcher, J.

(2018). RStan: The r interface to stan. r package version 2.17. 3.

Harris, K. M. (2013). The add health study: Design and accomplishments. *Chapel Hill: Carolina Population Center, University of North Carolina at Chapel Hill.*

Klugkist, I., Kato, B., & Hoijtink, H. (2005). Bayesian model selection using encompassing priors. *Statistica Neerlandica*, *59*(1), 57–69.

Kuh, D., & Shlomo, Y. B. (2004). *A life course approach to chronic disease epidemiology.* Oxford University Press.

Lynch, J., & Smith, G. D. (2005). A life course approach to chronic disease epidemiology. *Annu. Rev. Public Health*, *26*, 1–35.

Madathil, S., Joseph, L., Hardy, R., Rousseau, M.-C., & Nicolau, B. (2018). A bayesian approach to investigate life course hypotheses involving continuous exposures. *International Journal of Epidemiology*, *47*(5), 1623–1635.

Mulder, J., & Olsson-Collentine, A. (2019). Simple bayesian testing of scientific expectations in linear regression models. *Behavior Research Methods*, 1–14.

Mulder, J., Hoijtink, H., & Klugkist, I. (2010). Equality and inequality constrained multivariate linear models: Objective model selection using constrained posterior priors. *Journal of Statistical Planning and Inference*, *140*(4), 887–906.

Robert, C. (2007). *The bayesian choice: From decision-theoretic foundations to computational implementation.* Springer Science & Business Media.

Robins, J. M., Hernan, M. A., & Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. LWW.

Rosenthal, R., Rosnow, R. L., & others. (1985). *Contrast analysis: Focused* *comparisons in the analysis of variance*. CUP Archive.

Rosenthal, R., Rosnow, R. L., & Rubin, D. B. (2000). *Contrasts and effect sizes in behavioral research: A correlational approach*. Cambridge University Press.

VACEK, P. M. (1997). Assessing the effect of intensity when exposure varies over time. *Statistics in Medicine*, *16*(5), 505–513.

VanderWeele, T. J., Hernán, M. A., Tchetgen Tchetgen, E. J., & Robins, J. M. (2016). Re: Causality and causal inference in epidemiology: The need for a pluralistic approach. *International Journal of Epidemiology*, *45*(6), 2199–2200.

Table 1

Table 1. Confusion matrix for true and inferred life course models. Rows refer to the true models: i. accumulation; ii. linear sensitivity; iii. non-linear sensitivity; and iv. critical period. Columns refer to the inferred models: accumulation (a), sensitive (s) and critical (c) hypotheses, among simulations with evidence of a non-zero lifetime effect and inconclusive/unknown (u).

	a	s	c	u
i. a	14	0	0	4
ii. s	0	18	0	0
iii. s	0	17	0	1
iv. c	0	0	14	4

Table 2

Table 2. The finest credible ranking from a representative subset of simulations. Columns record the true model (ii or iii defined above), sample size, number of periods, true order and infered ranking (fcr), whether the infered ranking (fcr) violates or contradicts the true order, and the fraction of distinctions q that fcr preserves.

<table>
<tr><th>true_model</th><th>n_samples</th><th>n_periods</th><th>truth</th><th>fcr</th><th>violate</th><th>q</th></tr>
<tr><td>ii</td><td>700</td><td>3</td><td>1|2|3</td><td>1,2|3</td><td>FALSE</td><td>0.500</td></tr>
<tr><td>ii</td><td>3000</td><td>3</td><td>1|2|3</td><td>1|2|3</td><td>FALSE</td><td>1.000</td></tr>
<tr><td>iii</td><td>700</td><td>7</td><td>4,5,6,7|3|2|1</td><td>7,6,4,5,3|2|1</td><td>FALSE</td><td>0.667</td></tr>
<tr><td>iii</td><td>700</td><td>7</td><td>4,5,6,7|3|2|1</td><td>6,5,4,3,7,2|1</td><td>FALSE</td><td>0.333</td></tr>
<tr><td>iii</td><td>1500</td><td>3</td><td>3|2|1</td><td>3|2|1</td><td>FALSE</td><td>0.667</td></tr>
<tr><td>iii</td><td>3000</td><td>5</td><td>4,5|3|2|1</td><td>5,4|3|2|1</td><td>FALSE</td><td>1.000</td></tr>
<tr><td>ii</td><td>1500</td><td>7</td><td>1|2|3|4|5|6|7</td><td>1,2,3|4,5|6,7</td><td>FALSE</td><td>0.333</td></tr>
<tr><td>ii</td><td>1500</td><td>5</td><td>1|2|3|4|5</td><td>1|2,3,4|5</td><td>FALSE</td><td>0.500</td></tr>
<tr><td>iii</td><td>3000</td><td>3</td><td>3|2|1</td><td>3|2|1</td><td>FALSE</td><td>0.667</td></tr>
<tr><td>ii</td><td>700</td><td>3</td><td>1|2|3</td><td>1|2|3</td><td>FALSE</td><td>1.000</td></tr>
</table>

Table 3

Table 3. The mean proportion q of distinctions preserved by the posterior credible ranking increased with the simulated sample size.

n	q
700	0.52
1500	0.72
3000	0.89

Table 4

Table 4. The most probable partial ranking, at each level of granularity or informativeness.

Ranking	Probability
3|4|2|1|5	0.153
3,4|2|1|5	0.272
3,4,2|1|5	0.599
3,4,2,1|5	0.972
3,4,2,1,5	1.000

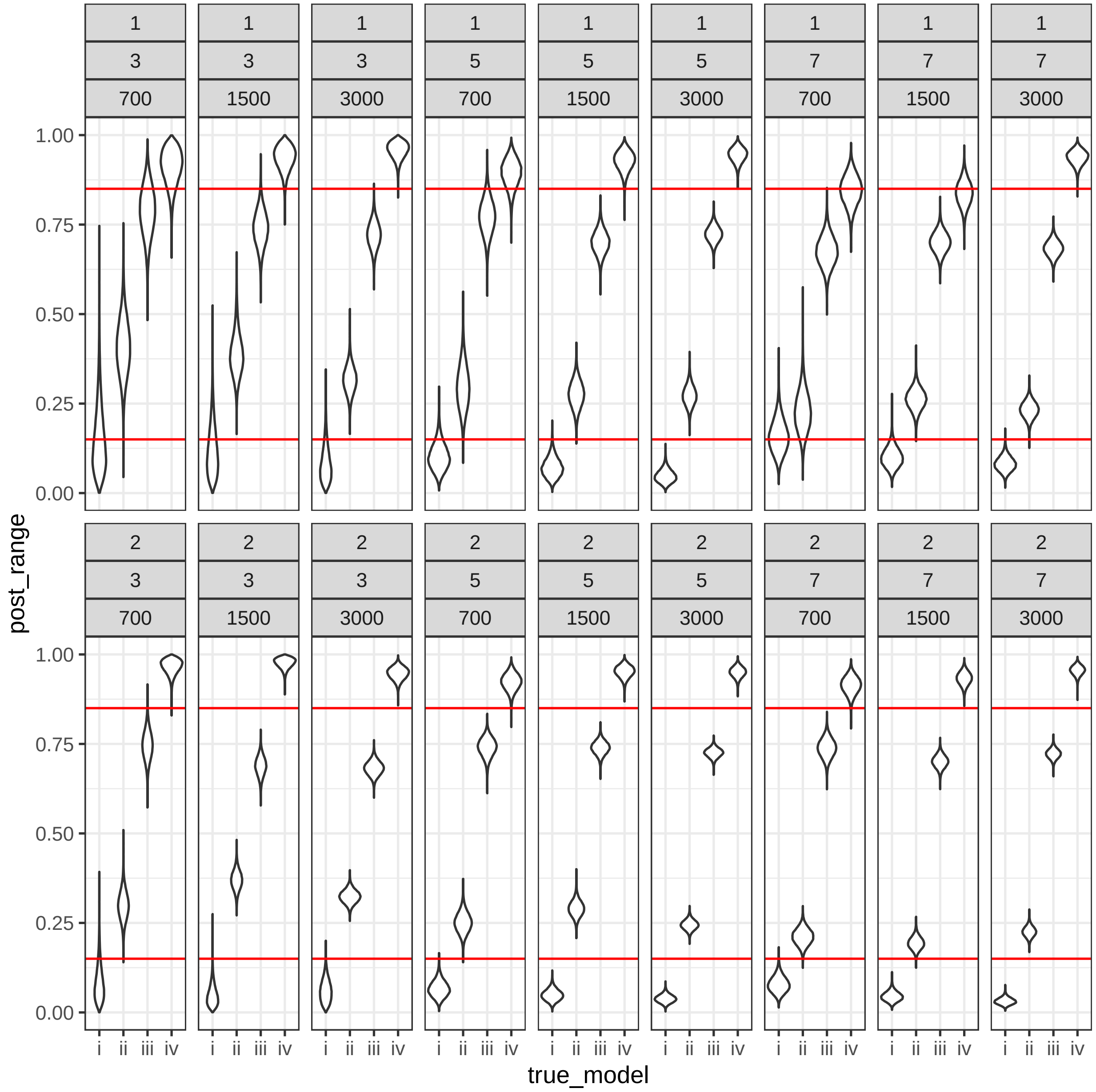

Figure 1. Figure 1. Posterior distribution of the range $\phi|y$ under different simulation conditions. Each cell is labeled (from top to bottom) with an integer giving the ground truth of the life-time effect δ^*, the number of periods T, and sample size n.

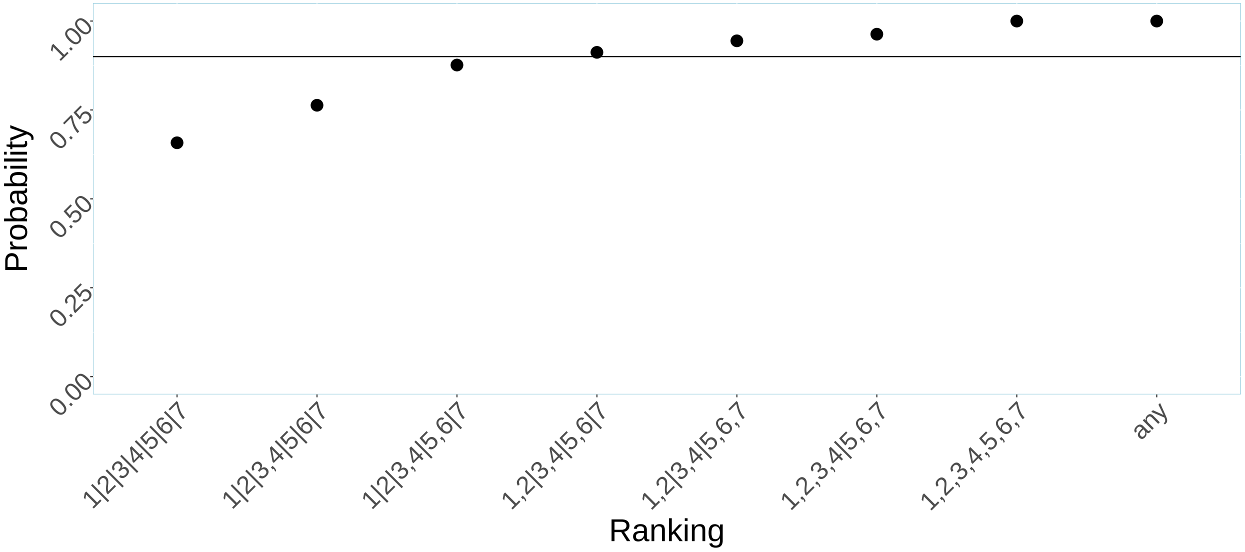

Figure 2. Figure 2. Posterior cumulative density over increasingly coarse partial rankings. The ground truth in this simulation was 1|2|3|4|5|6|7, an instance of model *ii* with monotonically increasing sensitivities. Progressing from left to right across the x axis rankings become coarser by the loss of one distinction ("|"). In this example, the partial ranking 1,2|3,4|5,6|7 is thus the finest ranking with 90% credibility. That is 1,2|3,4|5,6|7 is the 90% fcr: we conclude that measurement occasion 7 has greatest weight, occassions 5 and 6 (whose relative importance cannot be differentiated) have the next largest contribution to explaining outcome variation, then occassions 3 and 4 (again indiscriminable), and finally measurement occasions 1 and 2. Thus, in this example data-set, we can be confident about a coarse pattern of increasing sensitivity of outcome to exposures later in life. We do not however have high confidence that sensitivity is fully monotonically ordered: this conclusion warrants only ~70% credibility as given by the probability of the left-most entry on the x-axis, the full ranking 1|2|3|4|5|6|7).

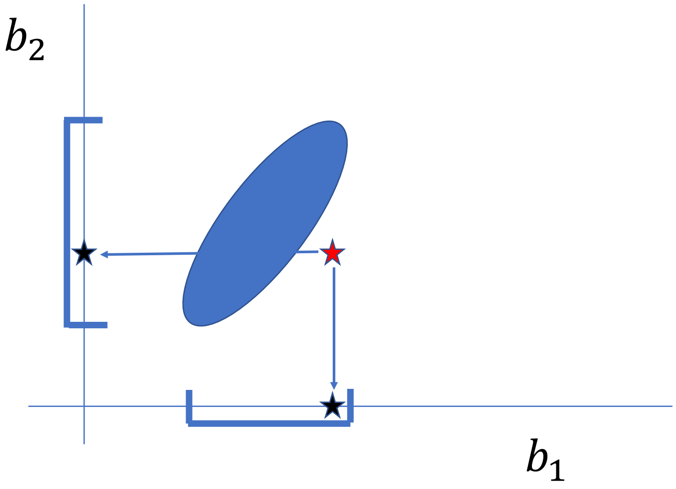

Figure 3. Supplementary Figure 1. One motivation for our proposed univariate transformation ϕ over multivariate confidence sets. Here we depict a 2 dimensional credible set for two arbitrary parameters (b_1, b_2) and its corresponding 2 univariate credible sets. The red star is a point in parameter space which is of theoretical interest. We might try to evaluate the "red star" theory by asking whether it is credible, i.e. whether it is within the credible set. In this illustration, the red star point hypothesis is *not* credible by this measure. Now consider the two 1 dimensional credible sets represented as intervals on the figure axes. These are much more convenient to calculate and interpret so it is tempting to use them for theory testing. They are derived simply by projecting or "marginalizing" the two dimensional joint posterior onto each component parameter. Similarly, the two black stars represent the joint hypothesis (red star), projected onto the margins. Crucially, both these blue stars now appear credible: they are *inside* both marginal credible intervals. This illustrates one hazard that arises with multidimensional credible sets. Our univariate method for model evaluation seeks to side-step this and other potential inconsistencies.