A Bayesian approach to comparing common models of life course epidemiology

Abstract

Background: Life course epidemiology studies people's health over long periods, treating repeated measures of their experiences (usually risk factors) as predictors or causes of subsequent morbidity and mortality. Three hypotheses or models often guide the analyst in assessing these sequential risks: the *accumulation* model (all measurement occasions are equally important for predicting the outcome), the *critical period* model (only one occasion is important), and the *sensitive periods* model (a catch-all model for any other pattern of temporal dependence).

Methods: We propose a Bayesian omnibus test of these three composite models, as well as post-hoc decompositions that identify their best respective sub-models. We test the approach via simulations, before presenting an empirical example that relates five sequential measurements of body weight to an RNAseq measure of colorectal cancer disposition.

Results: The approach correctly identifies the life course model under which the data were simulated. Our empirical cohort study indicated with >90% probability that colorectal cancer disposition reflected a sensitive process with current weight being most important but prior body weight also playing a role.

Conclusions: The Bayesian methods we present allow precise inferences about the probability of life course models given the data, and are applicable in realistic scenarios involving causal analysis, missing data.

Key messages

- Life course epidemiological methods often test or compare models via points in parameter space.
- Instead, we propose using *regions of practical equivalence* (ROPEs) for both model comparison and model decomposition.
- In particular, we describe the first general method for coherently decomposing the broad "sensitive model" into scientifically meaningful sub-models.

Introduction

Life course epidemiology typically examines associations between repeated measures of exposures and subsequent health over many decades of life. This approach has proven popular in the study of chronic forms of morbidity, which often involve multiple exposures over different developmental stages (hereafter "measurement occasions") and an appreciable latency period between exposure and outcome (1). Research in this area is generally guided by three life course models or hypotheses (2): the accumulation model, whereby multiple exposures of roughly equal importance predict an outcome; the critical period model, whereby one of multiple exposures (often very early in development) is decisively important; and the sensitive period model, whereby several exposures predict the outcome in non-trivial ways. Even though these models greatly simplify life course processes that are likely very complex, they have nonetheless proven influential and useful in the study of health. Adjudicating among these models—as well as several others that are relevant in certain cases (3) is thus a fundamental task in life course epidemiology.

Two main approaches have been proposed in the literature: a structured approach (4–6) and a Bayesian approach (7). The structured approach compares several nested models to a saturated model. Goodness of fit criteria are then used to ascertain the model that best fits the observed data. The Bayesian approach readily accommodates previous knowledge and arguably improves scientific interpretability (7–12), but is still in its infancy.

Madathil et al. (7) proposed certain Bayesian procedures for contrasting the critical, accumulation, and sensitive models given an observed data set. Their procedures are based on a non-linear transformation of the linear parameter $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_T)$ in a generalized linear model of outcome Y on the subject's exposure history $\mathbf{x} = (x_1, \ldots, x_T)$ over T periods. In particular, they estimate $\boldsymbol{\theta} = \delta\mathbf{w}$ where scalar δ captures "the total lifetime effect" and $\mathbf{w}$ is a set of *proportions* encoding the relative effect of each measurement occasion. Madathil et al. (7) then seek an alternative to conventional model selection by studying distance distributions and credible sets of posterior $p(\mathbf{w}|\mathrm{y})$. Some implicit assumptions of this approach are outlined in the Supplementary Material section.

In this paper, we introduce alternatives for summarizing the posterior distribution $p(\mathbf{w}|y)$ of Madathil et al. (7) and many related life course models. In particular, we compare the critical, accumulation, and sensitive models and sub-models by their posterior probability, rather than using complicated credible sets or the density of distances from competing point hypotheses (7) (see the base of Figure 1). Our approach is designed for situations of relatively high scientific uncertainty about which model is true. For example, without knowing which specific period may be critical in advance, we must examine all the critical models without succumbing to a multiple comparisons problem.

(FIGURE 1 HERE)

Unlike classical omnibus tests - e.g. F-tests or ANOVA – however we simply calculate the probability of three predefined regions of parameter space. Each region uniquely characterizes one model: it is the model's "region of practical equivalence" (ROPE). In detail, our proposed composite test involves first calculating the greatest difference between any two period's weights, that is, the range of component parameters $\mathbf{w} = (w_1, \ldots, w_T)$. This range characterizes the three models: it equals 0 and 1 for the

accumulation and critical models, respectively, and sits between these extremes for any sensitive model. In other words, the (univariate) range contains all the information required to construct ROPEs for the three broad life course models, see the base of Figure 1.

After assessing the composite models, post-hoc decompositions provide more detail. The critical model can very easily be decomposed by just examining the posterior components of w_i because exactly one will have high mass on high values, say $w_i > 0.85$. Crucially, we show that the sensitive model also decomposes nicely via a simple univariate construction, this time via *ranking* measurement occasions $1, \ldots, T$ by importance, *i.e.* which are more "sensitive" (have larger weights) than other time points. A simple algorithm then yields a concise conclusion such as "with 90% posterior credibility the first two measurement occasions matter more than any subsequent occasions". For reasons discussed below, we call this latter quantity the "finest credible rank". The final accumulation model requires no decomposition: there is only one way for all weights to be (approximately) equal.

While our goal here is not to improve causal identification *per se*, our proposed methods are general enough to be broadly applied in causal and non-causal frameworks. With or without causal identifiability, it is often important to rank parameters or to conclude that they are unrankable or tied (e.g., purely descriptive research). For simplicity, we therefore demonstrate our methods via the simple model of Madathil et al (2018) introduced above (where strong assumption of no feedback between time-varying treatments and time-varying confounders would be required to achieve causal identifiability) . In the discussion we explain how our methods apply under weaker identifiability conditions, non-linear models and models without a simplicial parameter space.

The paper is structured as follows: We first describe the mathematical rationale for our approach and then explain the goals and specifications of our simulation study. Having established the statistical validity of our approach, we apply our methods to an empirical example relating pronounced body weight over the life course to colorectal cancer disposition. The discussion then outlines the strengths and limitations of our approach.

Methods

Rationale

Figure 1 provides an overview of our proposed methodological strategy. Assuming the model of Madathil et al. (7), our generalized linear parameter is taken to be $\boldsymbol{\theta} = \delta \mathbf{w}$ where $\mathbf{w}$ is a vector whose components are proportions satisfying $w_t \geq 0, \sum_t w_t = 1$. Thus, $\mathbf{w}$ belongs to the T-part simplex, which we denote as Δ^T. Components of $\mathbf{w} = (w_1, \ldots, w_T)$ usefully capture the *relative* importance of exposure x at each measurement occasion $t = 1, \ldots, T$ for predicting outcome Y. Conversely, δ captures "the total lifetime effect."

We define our composite test by choosing two thresholds $a, b \in [0,1]$ to partition the values of ϕ into three intervals, which are practically equivalent to the $accumulation = [0, a]$, $sensitive = (a, b)$, and $critical = [b, 1]$ models. These intervals encode competing (non-

overlapping) models that are exhaustive: $p(sensitive\ or\ accumulation\ or\ critical|y) = 1$. We approximate the posterior probability of these three life course models by the fraction of posterior MCMC samples falling within the *accumulation*, *critical*, and *sensitive* intervals, respectively. We then conclude which model is most credible. For example, the sensitive model may be deemed credible in absolute terms if more than 90% of the posterior samples of $\phi|y$ fall within the *sensitive* interval. Or it may be deemed most credible in relative terms if the estimated Bayes factor

$$BF_{sensitive} = \frac{p(sensitive|y)/p(sensitive)}{p(critical\ or\ acculumulation|y)/p(critical\ or\ accumulation)}$$

exceeds some threshold.

The ROPEs (a, b) and $[b, 1]$ above encode "composite models". The posterior mass on $[b, 1]$ assesses the global critical model, i.e. is *any* of T critical periods plausible. Whenever this model is plausible, it is trivial to assess *which* component is critical: simply inspect the individual components of w because exactly one will markedly deviate from zero. Similarly, the sensitive model (a, b) encompasses *all possible full rankings*. Decomposing this latter is more tricky and leads us to now define the concept of "finest credible rank".

Let the function f label each point $\mathbf{w}$ with its full ranking. For example, f gives the point $\mathbf{w} = (0.2,0.7,0.1)$ the label 3|1|2 because the third measurement occasion is least important, followed by the first occasion, with occasion two being the most important. This notation for labelling full rankings is adapted from that of Lebanon and Mao (13). The full set of $T!$ full rankings comprise mutually exclusive sub-models of the sensitive model. We again approximate the posterior probability of these full rankings by the fraction of Markov Chain Monte Carlo (MCMC) samples satisfying the relevant inequalities ($w_3 < w_1 < w_2$ in this example). Thus, we can assess whether the most probable full ranking of measurement occasions by importance is, for example, 3|1|2, or whether $p('3|1|2'|y) \geq 0.90$. This assessment provides insights into our multivariate posterior $p(\mathbf{w}|y)$ without the inconvenience of constructing complicated continuous multivariate credible sets.

Importantly, the notation extends to coarser, *partial rankings*, such as 3,1|2, and calculate their posterior probability. A partial ranking can be viewed as a collection of full rankings: this ambiguity means they carry less information about the relative importance of measurement occasions. For example, the partial ranking 3,1|2 (equivalently denoted as 1,3|2) represents all points $\mathbf{w}$ that can be ranked as *either* $w_3 < w_1 < w_2$ or $w_1 < w_3 < w_2$: it, therefore, encodes points $\mathbf{w}$ for which w_2 is unambiguously the most important period. Then, if $p('1,3|2'|y) \geq 0.90$, we can say with 90% credibility that the second measurement occasion is most important, even though we can say nothing about the relative importance of the first and third occasions. The proposed algorithm starts at the maximum full ranking *a posteriori*, then recursively seeks the maximum *a posteriori* ranking at the next level of resolution (among all partial rankings with one additional bar "|" exchanged for a ","). This is depicted in Figure 1 and gives the optimal nested sequence of sub-models of the sensitive model. We define the $\beta\%$ "finest credible rank" (FCR), denoted as $\mathcal{C}_\beta$, as the first (finest, or most informative) such partial ranking with $\beta\%$ posterior probability. This is the most

informative statement that can confidently be made about the relative sensitivities of different measurement occasions.

(FIGURE 2 HERE)

Figure 2 illustrates this recursive scheme for whittling down the sensitive model using a hypothetical numerical example. It illustrates inference from data where we know the true full ranking was 1|2|3|4|5|6|7. The figure gives the "cumulative density function" of the best nested sequence of subsets of Δ^T. Among these, $\mathcal{C}_{90\%}$ is the finest 90% credible ranking (represented by the fourth point from the left in Figure 1). 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 this example, the partial ranking 1,2|3,4|5,6|7 is 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 the greatest weight, occasions 5 and 6 (whose relative importance cannot be differentiated) make the next largest contribution to explaining the outcome variation, followed by occasions 3 and 4 (again indistinguishable in terms of importance), and finally occasions 1 and 2. Thus, in this hypothetical example, we can be confident of a coarse pattern of increasing sensitivity of outcome to exposures later in life. However, we are not entirely confident 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.

As our discussion of Figure 2 illustrates, we require high posterior precision to faithfully identify the true full underlying ranking, such that our 90% finest credible rank is a *full* ranking. When posterior uncertainty is higher, we can only partially infer the parameter ranking, i.e. the 90% finest credible interval will be a *partial ranking* (possibly even the trivial ranking). This can also be pictured by contracting or expanding the isocontours in our schematic posterior density in Figure 1.

A detailed discussion of alternative priors on ϕ and f is presented in the Supplementary Material section.

Goals of the simulation

We ask whether the univariate credible interval for $\delta|y$ appropriately excluded zero, that is, whether it correctly inferred if *any* time period is relevant for the outcome. If the answer is positive, it makes sense to broadly examine the three competing models (critical, accumulation and sensitive) via the posterior distribution $\phi|y$. If $\phi|y$ additionally supports a sensitive model, $f|y$ and the probable ordering of weights should be examined. Therefore, the objectives of the simulation study were (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 models, 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, as discussed next.

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 as $\mathcal{C}_\beta$.

1) Is the FCR $\mathcal{C}_\beta$ consistent with the true full ranking? We say it is inconsistent if there is any $i \neq j$ for which the FCR asserts $w_i < w_j$ while in fact $w_j < w_i$ in the simulated ground truth. Otherwise, we say it is consistent.

2) How much "information" does $\mathcal{C}_\beta$ retain? Here, we use $q = r/r^*$ with values between 0 and 1 to measure the quality of information in $\mathcal{C}_\beta$, where r is the number of distinctions (inequalities or bars "|") in $\mathcal{C}_\beta$ and r^* is the true number in $f(\mathbf{w}^*)$. Therefore, a larger q 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 for $\mathcal{C}_\beta$ to be as informative as possible. Ideally, it should faithfully retain *all* distinctions made in the true ranking $f(\mathbf{w}^*)$.

Simulation parameters

Our simulation fully reproduced and extended that of Madathil et al. (7). 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 models (the two distinct sensitive models defined below, in addition to the accumulation and critical model) and three sample sizes ($n = 700, 1500, 3000$). The ground truth weight values of the simulation, denoted with an asterisk "$*$," were as follows: (i) pure accumulation model $\mathbf{w}_i^* = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$; (ii) linear 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 then extended the above three-period simulation to five and seven periods as follows: The accumulation model (i) above was generally $\mathbf{w}_i^* = (\frac{1}{T}, \ldots, \frac{1}{T})$; (ii) was generalized in the obvious way to $\mathbf{w}_{ii}^* = \frac{1}{\sum_{i=1}^{T} t}(1,2,\ldots,T)$, and (iii) and (iv) were padded with zeros, 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 seven dimensions.

We independently varied the lifetime effect δ^* between 0,1 and 2. The values simulate 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 as "$*$," to distinguish them from their posterior inferred counterparts. Given δ^* and $\mathbf{w}^*$, 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 $i = 1, \ldots, n$.

The prior distribution and inference

In accordance with the data-generating model above, we used Bayesian linear regression for inference. For the moment we follow Madathil et al. (7) 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 prior bias toward the critical model and $\alpha > 1$ implies a prior bias towards accumulation. We revisit these options below. We also chose a weakly informative Cauchy prior for the lifetime effect $p(\delta) = Cauchy(\delta|0,2.5)$.

To summarize, the model was

$$\begin{aligned} y_i &= \delta \sum_{j=1}^{T} x_{ij} w_j + \epsilon_i \\ \epsilon_i &\overset{i.i.d}{\sim} normal(0, \sigma) \\ \delta &\sim Cauchy(0, 2.5) \\ \mathbf{w} &\sim Dir\left(\frac{1}{D}, \ldots, \frac{1}{D}\right) \sim uniform(\Delta^T) \\ \sigma &\sim lognormal(1,1) \end{aligned}$$

The Hamiltonian MCMC sampler implemented in Stan (14) was used to acquire 50k marginal posterior samples of $\delta|y$ and $\mathbf{w}|y$. After performing 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 $\mathbf{w}|$.

Results of simulation

Posterior lifetime effect δ and range ϕ

In every simulation, the univariate credible interval for $\delta|y$ appropriately included/excluded zero. This pattern suggests that $\delta|y$ is a faithful omnibus measure of lifetime effect. In simulations with evidence of non-zero $\delta|y$, we used $\phi|y$ to adjudicate between the critical, accumulation, and sensitive models.

(TABLE 1 HERE)

By choosing two thresholds $a = 0.15$ and $b = 1 - a$ in the support of $\phi|y$ we constructed three intervals "practically equivalent" to the accumulation, critical period, and sensitive models, as discussed in the "Rationale" section. Table 1 reports the resulting confusion matrix, which relates the inferred models (column variable) to the underlying truth (row variable). We made a conclusive choice between these cases whenever one of them had greater than 0.9 posterior probability. Otherwise, our inference was considered inconclusive/unknown (u). There were $9/72$ inconclusive inferences. Inconclusive

inferences reflect posterior uncertainty, which, in turn, reflects either small effects or sample sizes.

However, all the conclusive results in our simulation study correctly identify the ground truth. In general, there were no incorrect inferences in our study, such as would arise if, for instance, we concluded that the accumulation model was true when in fact data were generated under the sensitive model. For example, of the 18 simulations where the accumulation model was true (Table 1, row 1), inference was correct in 14 cases and inconclusive in four cases. This table shows that around 90% (63/72) of inferences were both conclusive and correct.

The results of Table 1 were qualitatively similar when we instead set $(a, b) = (0.1,0.9)$ or $(a, b) = (0.2, 0.8)$. In complementary analyses, we also used a conventional 3.2 threshold for the (log) Bayes factors discussed in "Rationale" above (estimated by the ratio of prior to posterior samples in each ROPE). Results were again qualitatively similar, albeit with slightly inferior performance: there were still no misclassifications, but more inconclusive results (more false negative Bayes factors below the threshold 3.2).

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

When the sensitive model was credible (i.e. the 18+17 cases in column s of Table 1), it made sense to ask which periods were more or less sensitive. For these cases, we calculated the finest partial ranking $\mathcal{C}_\beta$ of parameters.

Regarding the goals of the simulation, we found that

1) the inferred finest partial ranking $\mathcal{C}_\beta$ never violated the ground truth; and

2) on average, over all simulations, about 0.71% of the distinctions were preserved. Table 2 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 increased to 72% and 89% with higher sample sizes. This pattern shows how increased posterior precision in $\mathbf{w}|y$ translates to increased precision of the inferred rank.

(TABLE 2 HERE)

(TABLE 3 HERE)

Empirical example

Data and regression model

The data are from the National Longitudinal Study of Adolescent to Adult Health (Add Health), a representative study of US adolescents in grades 7–12 during 1994–95 who were followed into adulthood over five waves of data collection (15). Specifically, the present study combines body weight data from a) parental report of birth weight, b) Waves I and II

(12–20 years), c) Wave III (18–27 years), d) Wave IV (25–33 years), and e) Wave V, Sample 1 (33–42 years).

During Wave V, RNAseq abundance data from peripheral blood samples were collected (currently, n=1132 samples collected in the 2016-2017 window have been fully preprocessed; details of data collection and preprocessing can be found in the Supplementary Material). Our outcome variable was a scalar mRNA colorectal cancer signature constructed as the normalized, weighted mean of 127 up-regulated genes (given positive weights) and 2 down-regulated genes (given negative weights). These genes are collectively implicated in colorectal cancer biology (16). Normalization was performed using a reference-gene normalization procedure (17) that converts raw counts onto a log scale. We had both mRNA and phenotype data for 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 $\overline{y}_i$. Our models controlled for biological sex, race/ethnicity, age at the time of the survey during Wave 1, preterm birth status, 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.

The goal was to predict a pre-symptomatic mRNA colorectal cancer signature from five measurements of pronounced body weight, taken on five separate measurement occasions: pronounced birth weight (defined as a birth weight higher than 8.8 pounds or lower than 5.5 pounds) and then pronounced body weight (BMI $\geq$ 30) at Waves I-V. We wanted to know whether one measurement occasion was *critical* for predicting variation in this cancer signature, if all occasions mattered equally (*accumulation*), or whether multiple occasions mattered to different degrees (*sensitive*). In the last case, we sought a credible and informative statement about the relative importance of the measurement occasions.

A Bayesian MCMC approach was used to estimate the reparameterized regression model of Madathil et al. (7) as $\overline{y}_i = \alpha + \delta \sum_{i=1}^{5} w_i\, x_i + \mathbf{c}'\gamma + \epsilon_i$. Here, α is an intercept, δ and $\mathbf{w} = (w_1, \ldots, w_5)$ are as defined throughout this work, and γ is the regression parameter vector for our control variables. The priors were the same as for the simulations above, with the addition that $\gamma_i \sim Cauchy(0,2.5)$ independently for each control variable. Thus, assuming a uniform prior distribution on $\mathbf{w}$ and weakly informative normal prior distributions on δ and γ, MCMC was used to collect 50k samples from their posterior distributions using the R package RStan, as implemented by Guo et al. (18). In particular, we pooled post-burn in samples from five independent chains of 10k after performing standard diagnostics to ensure convergence (we ran five independent parallel Hamiltonian Monte Carlo chains, each with 20k iterations and convergence was confirmed by checking trace plots and Rhat values).

Results

The posterior mean weight vector was w = (0.174, 0.104, 0.055, 0.077, 0.590). The posterior mean δ was 0.14 with 95% credible interval (0.09, 0.19). The central 90% and 95% posterior intervals for the lifetime effect δ were [0.1, 0.18] and [0.09, 0.19], respectively. These exclude zero, so we proceeded to evaluate the suitability of the accumulation, critical, and sensitive

models. We did this by choosing two thresholds $a = 0.15, b = 0.85$ to partition the posterior $\phi|y$ into three intervals (recalling that ϕ is defined as the difference between the largest and smallest components of $\mathbf{w}$.)

The posterior probabilities for the accumulation, sensitive, and critical models were 0.0004, 0.9973, and 0.0023, respectively. The sensitive model was thus the most credible, warranting an examination of the relative importance of the measurement occasions by examining the probable ranking of weights. Table 3 presents the highest probability partial rankings at each level of granularity. The most probable *full* ranking was 3|4|2|1|5. However, this highly informative statement carried only 19.7% posterior credibility. In fact, the 90% finest credible ranking, with 94.1% credibility, was 3,4,2,1|5, which can equally be written as 1,2,3,4|5 in our notation (because, in our notation, the order of numbers *between* any two adjacent bars is arbitrary). Thus, the final (fifth) measurement occasion matters more than any other (birth or Waves I-IV) in predicting the outcome. We have already concluded that (some of) these earlier periods probably matter: the posterior probability that they all have a weight of zero (i.e. the probability of a *critical* pattern of weights) is only 0.0023. 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. It should also be noted that with 74% credibility Wave 1 matters more for colorectal disposition than Waves 2–4, but less than Wave 5.

Discussion

Our approach builds on Madathil et. al.'s (7) proposal to use a Bayesian framework to evaluate different models of life course epidemiology. Like those authors, we pursue a detailed posterior analysis of a single model under weakly informative priors. We propose Bayesian composite tests and decomposition procedures based on the posterior probability of the critical, accumulation, and sensitive models and sub-models. Importantly, our approach relaxes dependence on continuous multivariate confidence sets, distance densities, and the requirement to specify point hypotheses ("expected weight vectors") (7). Instead, each model is identified with a region or "ROPE" of parameter space Δ^T, and we simply evaluate the posterior probability of that region. Our two-fold approach therefore simplifies the comparison of common models of life course epidemiology.

Our simulations illustrated that our methods were not capricious: adjudicating between the composite life course models never pointed to the wrong model, and model decomposition rarely confused the relative importance of measurement occasions. Therefore, we studied an empirical data example examining associations between an mRNA-based cancer signature and body weight history using Add Health cohort data. Our empirical example was motivated by the fact that high BMI is associated with a high risk of colorectal cancer (19, 20), which is among the most common cancers and a leading cause of cancer death (21). However, the role of BMI at different points in the life course has not been examined. Our method concluded that the critical and accumulation models do not offer a compelling explanation for the data. Instead, a sensitive pattern emerged in which contemporary pronounced body weight status was most salient to the abundance levels of

mRNA for genes associated with colorectal cancer and pronounced body weight at earlier ages (and possibly high/low birth weight status) was also independently predictive. Such findings, combined with prior studies (22), support the need for further study with larger samples that include data on life course patterns of body mass.

Our proposed methodological approach may be particularly attractive in situations of high prior uncertainty about the true relative importance of different measurement occasions, which is often the case in life course epidemiology. Assuming the sensitive model (or critical model) is true, there may be no compelling reason to hypothesize one particular pattern of temporal sensitivity to exposure over another. In such cases, it often makes sense to place a uniform distribution over parameter space Δ^T and, therefore, over different patterns of sensitivity or criticality. Our sensitive decomposition relies on the (component-wise) *order* of parameters. A precursor of this idea appears in Table 2 of Madathil et al (25), albeit on an ad-hoc subset of sensitive models (see also Madathil et al 26). Our work formalizes this idea into an algorithm which explicitly optimizes over the set of all possible decompositions (rankings), while controlling multiple comparisons. We have achieved this by rigorously defining and assessing the error, and information content, of an inferred decomposition, see "Goals of the simulation". The result is a method which complements other common summaries of the posterior distribution (e.g., mean and median).

Our methods have some limitations. First, our approach is unsuitable when exposures are measured at many, closely spaced time points (27). However, alternatives are suggested by Madathil et al. (7). Second, the posterior precision of our methods requires relatively large sample sizes (as discussed in the Results section) and, therefore, may be best suited to large epidemiological panels. Third, our simple expository model is too restrictive for many applications. For example, it is essentially linear in the parameters and assumes without justification that all coefficients have the same sign. Neither can its parameters benefit from a causal interpretation except under quite strong assumptions (28, 29). Fourth, assuming we are happy with the broad geometry of our ROPES – see Supplementary material for a critique and discussion – there remains the question of how to set threshold values a, b to define our composite tests. In what follows we discuss and seek to resolve these last two concerns.

Despite the attractive parameterization we have adopted from Madathil et al (7), their model is not explicitly causal or longitudinal: it has the same limitations and explanatory power as a generalized linear model on the restricted parameter space $\mathbb{R}_{\mp}^{T} = \pm(0,\infty)^T \subseteq \mathbb{R}^T$. In this restricted space $\mathbb{R}_{\mp}^{T}$, all components of any parameter vector must have the same sign (i.e., each measurement occasion affects the outcome in the same direction). For example, in the 2-dimensional case $\mathbb{R}_{\mp}^{2}$ this region is simply the union of the positive and negative quadrants. But we can easily promote this untested assumption to an explicit model, with posterior probability given by $p(\mathbb{R}_{\mp}^{T}|y)$ under a standard, *unrestricted* or "encompassing" Bayesian generalized linear model with parameters anywhere in $\mathbb{R}^T$. Namely, with a lower computational cost we can simply estimate an unrestricted Bayesian generalized linear model and assess the proportion of posterior samples in $\mathbb{R}_{\mp}^{T}$.

Moreover, we can also evaluate the two halves of this union separately: so $p(\mathbb{R}_+^T|y)$ for example examines the more specific claim that all components are positive, i.e. $\delta > 0$ in the parameterization of Madathil et al (7). Given posterior evidence for one quadrant (e.g., $\mathbb{R}_+^T$), a simple transformation $\theta_i \mapsto \theta_i / \sum_{i=1}^T \theta_i$ from this quadrant takes us back to the simplex Δ^D where all our proposed methods again apply. This observation indicates that our proposed methods apply in any model with a parameter vector, not the model of Madathil et al (2018). In fact, our procedure for identifying the FCR does not itself require the transformation $\theta_i \mapsto \theta_i / \sum_{i=1}^T \theta_i$ and can be applied directly to the encompassing vector space $\mathbb{R}^T$. Similarly, $\phi = 0$ characterizes the equality or accumulation model in both $\mathbb{R}^T$ or Δ^T, depending on which is most convenient. The accumulation model can therefore always be assessed by measuring the posterior mass of $\phi|y$ within a ROPE around zero.

In summary, our methods apply generically in the sense of not requiring many modeling assumptions, such as parameter linearity, error independence (e.g. temporal correlations), or strong causal identification conditions. They consequently apply to lists of real-valued parameters in one's favorite epidemiological model, to variance components of a multilevel model, etc. For example, the parametric g-formula is amenable to a Bayesian approach, e.g. Keil et al (30). Furthermore, our methods can be combined with any procedure that addresses missing data, so long as this procedure returns a well-defined posterior distribution. For example, one may mix draws from the posterior distributions of each multiply-imputed (complete) dataset (e.g. Zhou and Reiter 31).

We now consider how to set threshold values a, b. This specification is important partly because it affects the prior probability of our three composite models. One might adjust for this prior influence via relative measures such as $p(sensitive|y)/p(sensitive)$ or the Bayes factors discussed in our "Rationale". But these statistical solutions do not free us from considering the scientific meaning of our models. The generic – and unhelpful - guidance is to ensure the model ROPEs capture the scientific meaning of "practical equivalence" in one's particular application. Recall that ϕ is the size of the interval containing all components of the underlying parameter. Recall also that the accumulation model means all components cluster in a "small" interval around $(\frac{1}{T}, \dots, \frac{1}{T})$. We are therefore required to set threshold a "small" enough to match the scientifically appropriate notion of "small". Similarly, ϕ is highest when one period dominates all the others. Hence, we must set b "large enough" so that one period must dominate the rest in a scientifically "critical" fashion. In other words, we cannot avoid assigning precise meanings to the otherwise descriptive terms "critical," "accumulation," and "sensitive", although such meanings can themselves be subject to sensitivity analyses.

Furthermore, we discourage generically specifying a, b to ensure a uniform distribution over the three competing ROPEs: this speciously puts the three models on equal footing but poorly operationalizes the life-course models[1]. In situations of very high scientific

[1] Given our uniform distribution on weights $p(\mathbf{w})$, it can easily be shown that this definition undermines the interpretation of ROPEs as "critical", and "accumulation" (the ROPEs are too large). For example, this specification fails to guarantee that the critical ROPE contains only points for which a *unique* component *i* is high. Similarly, the accumulation ROPE may contain points far from the canonical accumulation vector

uncertainty, one palatable *generic* specification is $(a, b) = (0,1)$, i.e., we accept the sensitive model without question. This is appropriate when we are truly ignorant about the parameters: without auxiliary ROPEs, a uniform distribution on weights implies we believe the sensitive model[2]. In this situation we simply proceed to independently identify the FCR. Our simulations take an intermediary position. Our choice of $(a, b) = (0.15,0.85)$ somewhat boosts the critical and accumulation models while preserving the natural advantage of the sensitive model. As a consequence, we placed a higher bar on posterior support for the accumulation and critical models. Conversely, because of our prior bias for the composite sensitive model, it should be accompanied by its (unbiased) FCR model decomposition.

Despite these limitations, our proposed framework allows researchers to test central models of life course epidemiology. We presented straight-forward extensions to address some often-encountered complications in life course modeling: e.g., causal modelling, missing data, and situations involving auto-correlated errors (or nested data). The framework requires decisions to quantify the ROPES (a, b) and also criterion by which to draw substantive conclusions from the probabilities of the rankings (Figure 2) – specifications which will put tests of life course models on firmer scientific grounds.

Acknowledgements

We would like to thank Margaret Bellamy.

Funding

This research was supported by R01-HD087061 and the Jacobs Center for Productive Youth Development at the University of Zürich. We use data from Add Health, a project directed by Kathleen Mullan Harris and designed by J. Richard Udry, Peter S. Bearman, and Kathleen Mullan Harris at the University of North Carolina at Chapel Hill and funded by grant P01-HD31921 from the Eunice Kennedy Shriver National Institute of Child Health and Human Development, with cooperative funding from 23 other federal agencies and foundations.

$(\frac{1}{T}, .., \frac{1}{T})$. To upweight the accumulation and critical models without increasing the volume of their ROPEs, we would need the prior $p(\mathbf{w})$ itself to be non-uniform (recall that we can always set the concentration parameter of the Dirichlet to favor the critical or accumulation models, if this is desirable).

[2] This is because overwhelmingly comprises vectors whose components can be fully ranked, and collectively these "sensitive" vectors have prior probability 1. The posterior sensitive ROPE in this setting is itself meaningless: it vacuously has posterior probability 1 because it was given prior probability 1. Any interesting inference then hangs on whether the FCR is or is not itself the vacuous rank $1,2, .., T$.

Conflict of interest

None declared

REFERENCES

1. Lynch J, Smith GD. A life course approach to chronic disease epidemiology. Annu Rev Public Health. 2005;26:1–35.

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

3. Kuh D, Shlomo YB. A life course approach to chronic disease epidemiology. Oxford: Oxford University Press, 2004.

4. Mishra G, Nitsch D, Black S, De Stavola B, Kuh D, Hardy R. A structured approach to modelling the effects of binary exposure variables over the life course. International Journal of Epidemiology. 2009;38(2):528–37.

5. Smith AD, Heron J, Mishra G, Gilthorpe MS, Ben-Shlomo Y, Tilling K. Model selection of the effect of binary exposures over the life course. Epidemiology. 2015;26(5):719.

6. Smith AD, Hardy R, Heron J, Joinson CJ, Lawlor DA, Macdonald-Wallis C, et al. A structured approach to hypotheses involving continuous exposures over the life course. International Journal of Epidemiology. 2016;45(4):1271–9.

7. Madathil S, Joseph L, Hardy R, Rousseau M-C, Nicolau B. A Bayesian approach to investigate life course hypotheses involving continuous exposures. International journal of epidemiology. 2018;47(5):1623–35.

8. Greenland S. Bayesian perspectives for epidemiological research: I. Foundations and basic methods. International Journal of Epidemiology. 2006;35(3):765–75.

9. Greenland S. Bayesian perspectives for epidemiological research. II. Regression analysis. International Journal of Epidemiology. 2007;36(1):195–202.

10. Etzioni RD, Kadane JB. Bayesian statistical methods in public health and medicine. Annual Review of Public Health. 1995;16:23–41.

11. MacLehose RF, Hamra GB. Applications of Bayesian methods to epidemiologic research. Current Epidemiology Reports. 2014;1(3):103–9.

12. Robert C. The Bayesian choice: From decision-theoretic foundations to computational implementation. New York; Springer Science & Business Media, 2007.

13. Lebanon G, Mao Y. Non-parametric modeling of partially ranked data. Journal of Machine Learning Research. 2008;9:2401–29.

14. Carpenter B, Gelman A, Hoffman MD, Lee D, Goodrich B, Betancourt M, et al. Stan: A probabilistic programming language. Journal of Statistical Software. 2017;76(1).

15. Harris KM, Halpern CT, Whitsel EA, Hussey JM, Killeya-Jones LA, Tabor J, Dean SC. Cohort profile: The national longitudinal study of adolescent to adult health (add health). International journal of epidemiology. 2019 Oct 1;48(5):1415-k.

16. Guinney J, Dienstmann R, Wang X, De Reyniès A, Schlicker A, Soneson C, et al. The consensus molecular subtypes of colorectal cancer. Nature Medicine. 2015;21(11):1350-1356.

17. Eisenberg E, Levanon EY. Human housekeeping genes, revisited. Trends in Genetics. 2013;29(10):569–74.

18. Guo J, Lee D, Sakrejda K, Gabry J, Goodrich B, De Guzman J, et al. RStan: The R interface to Stan. R package, version 2.17. 3. 2018.

19. Ma Y, Yang Y, Wang F, Zhang P, Shi C, Zou Y, et al. Obesity and risk of colorectal cancer: A systematic review of prospective studies. PLOS One. 2013;8(1).

20. Campbell PT, Newton CC, Newcomb PA, Phipps AI, Ahnen DJ, Baron JA, et al. Association between body mass index and mortality for colorectal cancer survivors: Overall and by tumor molecular phenotype. Cancer Epidemiology Biomarkers and Prevention. 2015;24(8):1229–38.

21. Siegel R, DeSantis C, Jemal A. Colorectal cancer statistics, 2014. CA: A Cancer Journal for Clinicians. 2014;64(2):104–17.

22. Spracklen CN, Wallace RB, Sealy-Jefferson S, Robinson JG, Freudenheim JL, Wellons MF, et al. Birth weight and subsequent risk of cancer. Cancer Epidemiology. 2014;38(5):538–43.

23. Jensen BW, Gamborg M, Gögenur I, Renehan AG, Sørensen TI, Baker JL. Childhood body mass index and height in relation to site-specific risks of colorectal cancers in adult life. European Journal of Epidemiology. 2017;32(12):1097–106.

24. Smith NR, Jensen BW, Zimmermann E, Gamborg M, Sørensen TI, Baker JL. Associations between birth weight and colon and rectal cancer risk in adulthood. Cancer Epidemiology. 2016;42:181–5.

25. Madathil S, Blaser C, Nicolau B, Richard H, Parent MÉ. Disadvantageous socioeconomic position at specific life periods may contribute to prostate cancer risk and aggressiveness. Frontiers in oncology. 2018 Nov 15; 8:515.

26. Madathil S, Rousseau MC, Joseph L, Coutlée F, Schlecht NF, Franco E, Nicolau B. Latency of tobacco smoking for head and neck cancer among HPV-positive and HPV-negative individuals. International Journal of Cancer. 2020 Jul 1;147(1):56-64.

27. Vacek PM. Assessing the effect of intensity when exposure varies over time. Statistics in Medicine. 1997;16(5):505–13.

28. Robins JM, Hernan MA, Brumback B. Marginal structural models and causal inference in epidemiology. Epidemiology; 2000;11(5):550-560.

29. VanderWeele TJ, Hernán MA, Tchetgen Tchetgen EJ, Robins JM. Re: Causality and causal inference in epidemiology: The need for a pluralistic approach. International Journal of Epidemiology. 2016;45(6):2199–200.

30. Keil AP, Daza EJ, Engel SM, Buckley JP, Edwards JK. A Bayesian approach to the g-formula. Statistical methods in medical research. 2018 Oct;27(10):3183-204.

31. Zhou X, Reiter JP. A note on Bayesian inference after multiple imputation. The American Statistician. 2010 May 1;64(2):159-63.

Figures

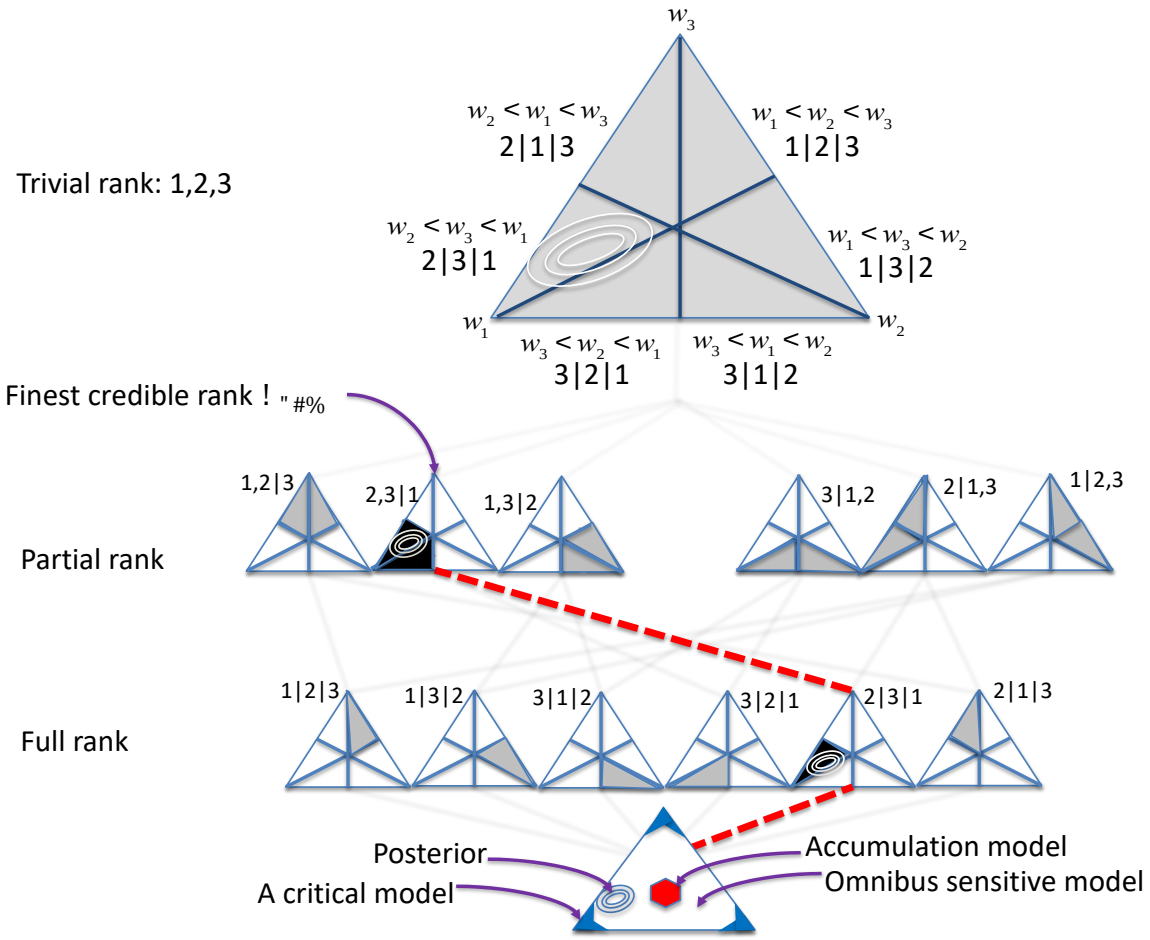


Figure 1. This graphic depicts our proposed method by considering the posterior distribution of a schematic three life-period model (for ease of presentation) on Δ^3. The diagram reads bottom to top. The schematic posterior probability iso-contours are depicted inside the simplex. Starting at the base, we first evaluate posterior mass on the omnibus sensitive, critical and accumulation models using ϕ. Specifically, we chose thresholds $(a, b) = (0.15, 0.85)$ which imply the polygonal ROPEs in the simplex at the base of the figure. The region shaded in red corresponds to the interval $[0, a] = [0, 0.15]$, blue region corresponds to the intervals $[b, 0] = [0.85, 1]$, and non-shaded region corresponds to the interval $(a, b) = (0.15, 0.85)$. (See Supplementary Material for a higher resolution image and critique.) Our schematic posterior is consistent with the sensitive model, motivating us to ask "*which* specific sensitive sub-model?" We answer this question by advancing up the diagram (choosing the most probable event at the next level up). Note that upward paths correspond to the subset inclusion relation $\subseteq$, so posterior probability monotonically increases accordingly: see Figure 2 for a numerical example. We stop at the first region with probability greater than 95% and call this region the 95% finest credible region. Our procedure is inconclusive if (and only if) the first such region is the vacuous ranking at the top of the diagram. Now suppose that, from the base of the diagram, our posterior had instead supported the critical model. Then our follow-up decomposition could identify

which critical period by simply examining $p(w_i > t|y)$ for some high threshold t. Finally, had we chosen the accumulation model at the base of the diagram, no further decomposition would have been required.

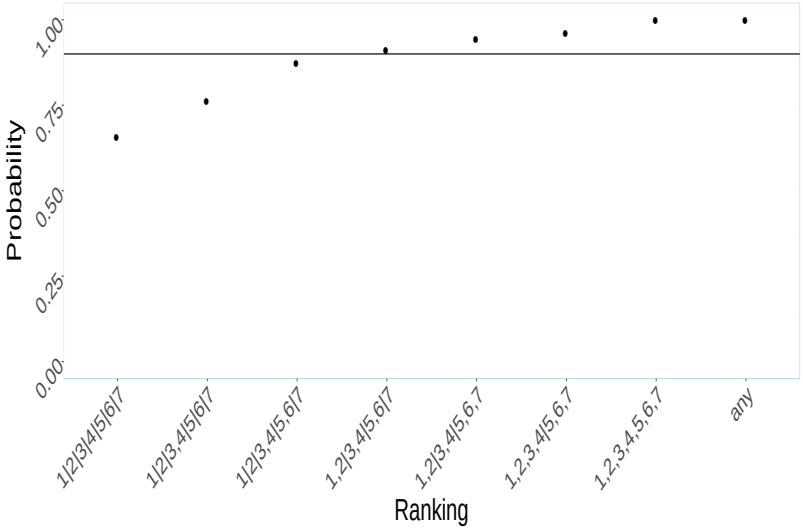


Figure 2. Posterior cumulative density over increasingly coarse partial rankings. The ground truth in this example was 1|2|3|4|5|6|7*. Progressing from left to right across the x-axis, rankings become coarser by the loss of one distinction ("|"). All points above the horizontal black line have at least 90% posterior credibility.*

Tables

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, as well as 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. Simulated sample sizes and mean proportions q of distinctions preserved by the posterior credible ranking. The quantity q refers to the proportion of distinctions in the ordering of the underlying parameter, which are preserved in the inferred ordering of that parameter, i.e. the FCR.

n	q
700	0.52
1500	0.72
3000	0.89

Table 3. Ranking measurement occasions by their importance for colorectal cancer (i.e. their relative magnitude). The best sequence of nested submodels (partial rankings) of the sensitive model and their posterior probability.

Ranking	Probability
3|4|2|1|5	0.197
3|4,2|1|5	0.350
3,4,2|1|5	0.737
3,4,2,1|5	0.941
3,4,2,1,5	1.000

A Bayesian approach to common models of life course epidemiology

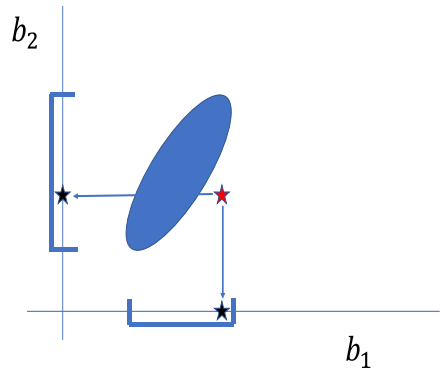
Supplementary material

Implicit assumptions in prior work

We elaborate on two assumptions required by Madathil et al.'s (1) approach. The first concerns the use of credible sets. To compare the three broad life course hypotheses, Madathil et al. determine whether the posterior 95% credible regions exclude or include (cover) the accumulation or critical hypotheses which are treated as points in parameter space. 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 $\mathbf{w} = (0,0,1)$. This approach is properly implemented using multivariate credible sets, but the limitations of continuous, multidimensional credible sets create challenges in bounded parameter spaces, such as Δ^T. 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 coverage implies joint coverage. In fact, a multivariate set S may exclude a point p, even while all of P's lower dimensional (marginal) projections include all of p's projections.

Supplementary Figure 1 provides some indication of the potential problems arising from the use of multivariate credible sets. We might try to evaluate the "red star" hypothesis by asking whether it is credible, that is, whether it is within the multivariate credible set (two-dimensional ellipse). In Figure

1, the red star point hypothesis is not credible by this measure. If we consider the two one-dimensional credible sets represented as intervals on the axes, they are derived as having projected or "marginalized" the two- dimensional joint posterior onto each component parameter. Crucially, both these blue stars, sitting on the *x*- and *y*-axis respectively, now appear credible: they are inside both marginal credible intervals. For this and other reasons, we would ideally avoid multivariate credible sets.



Supplementary Figure 1. Two-dimensional credible set for two arbitrary parameters (b_1, b_2) and its corresponding two univariate credible sets. The red star is a point in parameter space which is of theoretical interest. Similarly, the two black stars represent the joint hypothesis (red star) projected onto the margins. Such situations arise both in familiar Euclidian vector spaces, as well as in the probability simplex Δ^T.

Second, the use of the density of distances from hypothetical points in parameter space requires a choice of a point hypothesis (which, ideally, should be avoided) and a choice of metric on parameter space (e.g., the Euclidean metric as opposed to the Hilbert projective metric or the Aitchison metric on the simplex). Our sensitive model decomposition is based on order relations and does not require any metric.

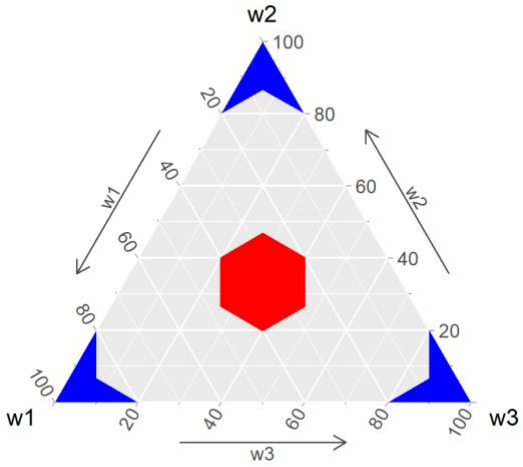
Defining alternative priors

Our first methodological proposal offers a new way of choosing between accumulation, critical, and sensitive models. In contrast to our proposed strategy and that of Madathil et al. (1), a more conventional approach to model selection using Bayes factors would require explicitly specifying point mass priors on the critical and accumulation points, as well as 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 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. (1). Like those authors, we choose a uniform prior over all Δ^T, but we proceed by 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 by, for example, setting the Dirichlet hyperparameter to $\alpha = 1$, actually

favours 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 are, therefore, 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 could introduce prior bias of the desired strength towards either the accumulation model ($\alpha > 1$) or the critical model ($\alpha < 1$).

Our second method decomposes the sensitive model, yielding an inference about which measurement occasions are more important for the outcome. It should be noted 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. In contrast to model adjudication via ϕ, the uniformity of $p(\mathbf{w})$ is, therefore, indeed preserved. In particular, each full ranking is uniformly 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 equalling $2/3!$ 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 (2–5).

Alternative ROPEs



Supplementary Figure 2: The preimages of [0, a], (a, b) and [b, 1] as polygons in the case of three time points. As the figure shows, our ROPES are not entirely free from criticism. For example, consider the thresholds of a = 0.2 and b = 0.8, the interval of ϕ that should be considered for the critical period hypothesis is [0.8, 1]. The ROPE here is a polygon with vertices at w = {(1.0, 0.0, 0.0), (0.8, 0.0, 0.2), (0.8,0.2,0.0), (0.8667,0.0667,0.0667)}. Interestingly, the point w = (0.85, 0.075, 0.075) is not in the ROPE (ϕ = 0.77); while the point w = (0.8, 0.2, 0.0) is in the ROPE. However, the former sample apparently supports a critical period hypothesis. As with any ROPE, readers must decide for themselves whether the geometry of these particular ROPES is acceptable. In simulations our ROPEs, which have a pleasing connection to the range or variability ϕ of weights, had satisfactory statistical performance.

While we have used ϕ to create ROPEs for omnibus models, there are many other reasonable and intuitive ways to create ROPEs. Such alternative constructions should be conceptually analogous to ours, differing only in their precise geometry. They may nonetheless give some intuition for our choice of ϕ. For example, rather than $\phi < a$, the accumulation ROPE could be taken as the set of $\mathbf{w}$ whose components are all within some small distance a' of $(\frac{1}{T}, .., \frac{1}{T})$. Similarly, rather than using $\phi > b$, the omnibus critical ROPE

could be taken as the set of $\mathbf{w}$ for which any component3 exceeds threshold b'. The sensitive omnibus ROPE is then just the complement of these two ROPEs, i.e. the rest of Δ^T. Like those based on ϕ, these alternative ROPEs can also be used when we are uncertain *a priori* about which specific critical model may be true. (They also offer a multiple comparisons adjustment.) These alternative ROPEs can fairly be compared with ours when (a', b') is chosen to satisfy

$$p(\phi < a) = p(\mathbf{w} \in \Delta^T : \forall i \; |w_i - 1/T| < a')$$

$$p(\phi > b) = P(\mathbf{w} \in \Delta^T : \exists i \;\; w_i > b').$$

Note that the latter has strictly greater probability than any one critical model i. This is because the monotonicity of probability gives $P(\mathbf{w} \in \Delta^T : \exists i \;\; w_i > b') \geq p(\mathbf{w} \in \Delta^T : w_i > b')$ where i is free in the former and fixed in the latter. It is obviously also smaller than $T \times p(\mathbf{w} \in \Delta^T : w_i > b')$. We numerically solved the two equations above in order to examine the behavior of these alternative ROPEs in our simulations. The ensuing posterior inferences were qualitatively alike, demonstrating some robustness to the exact geometry of the ROPEs.

mRNA preprocessing

FASTQ sequencing data was transformed into counts using the STAR aligner. ENSG gene identifiers were then mapped to a Gene Symbol and counts for the same gene symbol were summed (i.e., summing over alternative transcripts/versions of the same gene). Note that in many cases multiple distinct ENSG identifiers may be ascribed to a single gene symbol, and in other cases, there may be no gene symbol to map to. In the latter case, the ENSG identifier is retained.

Starting from 59068 gene ids, we removed haemoglobin genes and genes with no HUGO ID, leaving 55772. We then removed genes with insufficiently large counts to be retained in a statistical analysis, resulting in 8484. The latter was performed in edgeR, using filterByExp, using a filtering strategy described by Chen & Smyth (2016). Roughly speaking, this strategy keeps genes that have at least 10 reads in a worthwhile number of samples.

We reference-gene normalized the count data, yielding log Transcripts Per Million (TPM). This involves dividing by the mean count over the 11 housekeeping genes identified by Eisenberg and Levanon (2013) (2013, Table 1): C1orf43, CHMP2A, EMC7, GPI, PSMB2, PSMB4, RAB7A, REEP5, SNRPD3, VCP, VPS29. See (6) for more details.

REFERENCES

1. Madathil S, Joseph L, Hardy R, Rousseau M-C, Nicolau B. A Bayesian approach to investigate life course hypotheses involving continuous exposures. International Journal of Epidemiology. 2018;47(5):1623–35.

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

---

3 Strictly speaking a *unique* component $\exists! \, i \;\; w_i > b'$. Uniqueness preserves the interpretation as a critical hypothesis, and is ensured whenever $b' > 0.5$, i.e. high enough that multiple components of the same point $\mathbf{w}$ do not simultaneously exceed b'.

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

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

5. Mulder J, Olsson-Collentine A. Simple Bayesian testing of scientific expectations in linear regression models. Behavior Research Methods. 2019;1–14.

6. Cole, SW, Shanahan, MJ, Gaydosh, L., & Harris, KM. In press. Inflammatory and antiviral gene expression in Add Health: Molecular pathways to social disparities in disease emerge by young adulthood. *PNAS*