ADVANCED MODELLING STRATEGIES

Challenges and pitfalls in robust causal inference with observational data

PETER WG TENNANT
KELLYN F ARNOLD
LAURIE BERRIE
GEORGE TH ELLISON
MARK S GILTHORPE

ADVANCED MODELLING STRATEGIES: Challenges and pitfalls in robust causal inference with observational data

Published by: **Leeds Institute for Data Analytics (LIDA)**, 2017

ISBN 978-1-5272-1208-4

Leeds Institute for Data Analytics (LIDA)
University of Leeds
Worsley Building
Clarendon Way
Leeds
LS2 9JT
United Kingdom
http://lida.leeds.ac.uk/
lida@leeds.ac.uk

Advanced Modelling Strategies was sponsored by the *Society for Social Medicine* and hosted by the *Leeds Institute for Data Analytics (LIDA)*.

TABLE OF CONTENTS

1. PREDICTION VS. CAUSAL INFERENCE

Learning objectives

- Understand the distinction between model prediction and model inference
- Recognise the main biases that impact observational research

Multivariable linear regression modelling

You should already be familiar with the concepts of linear regression modelling. In its most basic form, we have a univariate (one outcome) univariable (one 'explanatory' variable or covariate) regression model with continuous outcome. This can be extended to accommodate:

- Multivariable (multiple covariates) models with either count, nominal or ordinal (most often logit or probit) outcomes;
- Time-to-event outcomes that are parametric (e.g. Weibull) or semi-parametric (e.g. Cox proportional hazards) with continuous or discrete (interval censored) times;
- Multivariate (i.e. multiple outcomes) models, with two or more outcomes analysed simultaneously (e.g. systolic and diastolic blood pressure).

There are more complex statistical regression models (e.g. structural equation models), but here we focus on generalised linear regression models and examine potential pitfalls that lie with their use (and abuse) in biomedical research, particularly for observational data.

Prediction vs. Inference

To properly understand the issues arising with model inference, it is important first to distinguish between prediction and (causal) inference. What follows is a summary of the key features of prediction and inference for linear regression models.

With model **prediction**, one is concerned with:

- Maximising the proportion of explained outcome variation, i.e. the greater the R^2 the better the predictive model. Whilst seeking a maximal R^2 is desirable, this can be very data-specific and may not be replicable from one dataset to another. Furthermore, if variance reduction is all we seek, then 'retro-dictors' are as good as 'pre-dictors', i.e. including consequences of our outcome in the regression model would be as good as including its predictors!
- Predictive models are often developed on one dataset (the training data) and evaluated on a different dataset (the test data); in practice, the training and test data may be a single dataset randomly split into two. Failure to evaluate predictive models renders their utility limited, and all too often an inadequate effort is made to 'select' models that are generalizable - greater emphasis being placed on model fit for a single dataset rather than model parsimony.
- The smallest subset of covariates that yields the largest R^2 is often preferred, since the model is then less prone to inconsistency across different datasets, especially if the covariates are prone to measurement error (recall: linear models assume that covariates are error-free).
- The specific choice of covariates is unimportant, apart from minimising collinearity (see next).
- Collinearity can be a problem for predictive models because it introduces a lack of precision, thereby yielding large standard errors in both the predicted outcome and covariate coefficients. The latter can interfere with procedures for covariate subset selection, making determination of a good model difficult.

- The greater the statistical significance of a covariate, the greater the argument that is often made for retaining it in the model; however, the basis of this can be questionable (see note below). The key point is that the size of the estimated covariate coefficient is not important *per se*; if statistical significance related to covariate coefficient size is considered, emphasis should be given to selection and retention of the most informative subset of covariates (i.e. in terms of accurately and precisely predicting the outcome).
- No causal interpretation of any model covariates should ever be sought; the predicted outcome is the sole focus of a predictive model.
- Although causal relationships are unimportant when selecting covariate subsets, to maximise the likelihood that the model derived for one dataset is also 'good' for other datasets, it can be helpful to select covariates that are deemed likely to have a causal relationship with the outcome. If there is a close call between two covariates where the procedure used to select variables slightly favours the covariate with a less obvious causal relationship to the outcome, overriding the procedure to select covariates in favour of those a potential causal link will likely improve generalisability of the predictive model.

Note: It has been proven that forward or backward **stepwise procedures** for variable selection (based on statistical significance when introducing or removing covariates) does not guarantee that an 'optimum' subset of covariates is obtained (i.e. one having the largest R^2). Stepwise procedures also typically explore covariates as if these were separate linear predictors, with no consideration given to nonlinear (curvilinear) relationships for each covariate or to potential interactions amongst all covariates. It will often require human intervention to settle what nonlinear relationships are required; while, with interactions, the model may become saturated (i.e. having more covariate combinations than observations) as well as being less parsimonious. Stepwise procedures are therefore inadequate for developing predictive models, and more appropriate methods to optimise these now exist[1]. It is interesting to note that many researchers (including statisticians) still use and advocate forwards/backwards step-wise procedures, instead of adopting the preferred methods now available – this is bad practice.

With model **inference**, the following is of concern:

- All putative **causal** relationships between covariates and outcomes (or the magnitude of any associations, if the direction of causality is unknown *a priori*) become the focus.
- The magnitude of associations between covariates and outcomes are examined (whilst mindful of potential confounding [see later]).
- Focus is on the size of specific covariate-outcome relationships, which may be inferred as **clinically significant** as opposed to statistically significant; this is described more generally in the literature as **'effect size'**.

Note: Inferential interpretation of a covariate's effect size is nearly always **context specific**, and this is not often appreciated (certainly not as much as it should be, even amongst experienced epidemiologists and biostatisticians). This is because each covariate-outcome relationship is dependent (i.e. conditional) on the **adjustment set** of all other covariates in the model. Effect sizes are always dependent upon the selected adjustment set and overlooking this can be misleading (it leads to what is known as the "Table 2 fallacy", which is covered in detail later). As different studies adjust for different sets of covariates, effect sizes between observational studies

may not be directly comparable (and it therefore requires great care when undertaking a synthesis/meta-analysis of observational studies).

The role of linear models in epidemiology

Although epidemiology is concerned with producing descriptive information (such as incidence and prevalence rates, or profiles and population trends), statistical modelling in epidemiology is primarily concerned with causal inference and less with prediction (albeit with notable exceptions, e.g. predicting outbreaks). To that end, we focus on statistical modelling in its approach to derive causal inference. First, it is helpful to reflect upon how and why epidemiology has gone about the various applications of linear modelling in the past, though perhaps not always drawing upon (or appreciating fully) the distinction between prediction and inference. We then look at how much has changed, quite recently, to introduce the basic principles of causal inference and associated techniques (such as causal graphical models). Finally, we explain how these developments impact on our current use of linear modelling.

Randomisation and the role of study design

The design of a *randomised* control trial (RCT) aims to ensure that differences between treatment groups are due to the causal effect of the treatment of interest. An RCT uses **randomisation** to balance the groups, such that each treatment arm is similar in every respect (i.e. regarding potential biases), apart from the treatment under study. In epidemiological investigations, an exposure may be of interest as having a putatively causal impact; this makes investigation of an exposure similar to the evaluation of treatment effects in an RCT, though without the involvement of randomisation. To overcome this limitation in the observational setting (i.e. the absence of groups being balanced via randomisation), epidemiological studies have relied on a 'top-down' approach of careful study design to address potential study bias.

For instance, **case-control studies** 'match' pairs of observations with the intention of creating balance across exposure groups akin to an RCT. This may be effective, depending upon the context; though many potential biases may remain. Another approach is the **cohort study**, which collects information prospectively, agnostic to the outcome, and thereby minimises the potential for biases due to what is termed 'confounding' (an issue we will formally define later).

It is debatable how effective case-control and cohort study designs are at eliminating all sources of bias, though it is interesting to note that case-control studies typically estimate a larger effect-size for the same exposure than do cohort studies. If the estimated effect is confirmed by an RCT, it is nearly always much smaller than that for either case-control or cohort studies, suggesting that both designs suffer biases (though perhaps cohort studies are superior to case-control studies, as they usually exhibit less bias associated with an inflated estimate of the causal effect). We do not dwell on the specifics here but note that neither study design is perfect in eradicating all potential biases in the way that randomisation can within a well-conducted (and appropriately randomised) RCT.

A theoretical framework for causal inference

The intention behind an RCT or epidemiological study is typically to uncover a putatively causal effect of treatment or exposure. A limitation throughout the history of epidemiology, however, has been the lack of a strong theoretical framework for causal inference in the absence of

randomisation … until relatively recently, at least. Attempts to embrace causal inference in epidemiology have therefore operated mainly through the elimination or reduction of confounder bias via study design (i.e. through the use of RCTs rather than observational data). A root problem in the absence of randomisation is that there are too many *knowns* and *unknowns* impacting upon the exposure and outcome. 'Controlling' for all potential confounders, either by study design or statistical sophistication (or both), has proven challenging. In part, this has been because methods which aid this, such as linear modelling, were beyond reach until the advent of personal computers towards the end of the last century. However, while linear models are now ubiquitous they do not, in and of themselves, address confounding. This still needs the formalisation of underpinning causal theory. It remains that linear models cannot address the issues of differential selection and errors in covariates; the latter requires more sophisticated modelling techniques, such as structural equation modelling (SEM).

Heavily utilised in the social sciences, SEM is less commonly encountered in epidemiology, but it can readily model covariate errors and has always sought to inform causal inference. However, the formal theoretical framework underpinning causal inference in SEMs has developed only recently to become robust; the firming up of ideas for more complex empirical contexts, such as differential selection, remains a work in progress. Some would describe recent developments as a causal inference 'revolution', though largely taking place in the computer sciences than in epidemiology. In effect, theoretical developments are crossing discipline boundaries very slowly. Perhaps a challenge to its widespread adoption is that the underpinning philosophy of causal inference theory is contrary to epidemiological convention. This new theoretical framework is 'bottom-up' in that it emphasises understanding of 'real-world' relationships and builds on these to obtain meaningful causal inference, with focus on the implications of what is observed. There is also no longer any reliance on study design to control data generation.

Effect size

It is somewhat unfortunate that the term 'effect size' has become entrenched in the literature on this topic, since effect implies causation. It may well be the case that a researcher wishes to test a causal hypothesis. However, it may instead be the case, for instance in a case-control study, that causation cannot be inferred. A more appealing term would be 'association size', but this is not used in the literature, so we will continue the discussion of 'effect size' with this caveat in mind. However, often we do mean to refer to causation (even if subliminally), though the reason we won't admit this to ourselves is because we are aware of the limitations of making such a claim in the absence of a robust causal theoretic framework.

It is important to realise that effect size is not a panacea. It is still a **statistic**, and since it is derived from a sample, it is a **random variable** with sampling variation and is therefore only an **estimate** of the 'true' effect size. It should always be reported together with a measure of uncertainty, such as a confidence interval.

Most importantly, effect size is not only affected by sample size and natural variation, but also by measurement error, differential selection, confounding and inappropriate statistical adjustment. These all distort the derived effect size such that – even when its estimate is mathematically correct when derived for a given study sample – its meaning may be substantially biased from the 'true' **causal interpretation** we give to the target population. Bias from error, differential selection,

and/or confounding may drown out natural sampling variation and render the meaning of a p-value (derived from effect size accounting only for sample variation) as extremely limited. Attention must also be given to dealing robustly with measurement error, differential selection, and confounder adjustment – all within a robust causal theoretical framework.

Biases in epidemiological studies

At this point it should be noted that we have not stated what biases might be operating. There are many descriptions of these issues in standard epidemiological texts, but for simplicity and ease of reference we refer to 3 types of bias in epidemiological studies: **error**, **selection** & **confounding**.

1. **Error** – the primary culprit is **measurement error**, typically associated with the instruments used to record data, though potentially caused by incorrect or inadequate application of the instruments. Another equally important – but often overlooked or conflated – source of *error* is within-subject **biological heterogeneity** (i.e. individual biological variations).

Errors can arise due to limitations in study techniques (e.g. where data retrieval processes go wrong) or bad coding practices (e.g. misclassification occurring where coding criteria are applied incorrectly). **Biological variation** arises due to volatility in the clinical measure sought, where the value of any one measure may vary around an underlying value that is representative of the overall 'state' (e.g. blood pressure may vary due to physical situation, such as resting vs. exercising, and is affected by context, such as 'white coat hypertension'[2]); this makes the assessment of a 'health state' challenging and prone to 'error', though it is not the measurement instrument that is at fault. Both aspects of error (**instrument error** and **biological variation**) are typically conflated as one source of error within most statistical techniques, including standard linear modelling.

2. **Selection** – as well as inherent sampling heterogeneity (i.e. the underlying statistical process of sampling), other factors may operate that limit the chances of capturing the relevant information either at all, or representatively. Subsequent statistical evaluation is mathematically correct but may yield estimates that do not generalise to the target population of interest.

Simple sample selection can be addressed by statistical methodological rigour, but other forms of selection bias may occur, as with missing or incomplete information, where complete cases go unrecorded. This results in an analytical sample that does not represent the target population. Selection bias may operate via **differential participation** (e.g. disadvantaged individuals tend to be less inclined/able to participate in case-control studies due to various life constraints, yet this diminished inclination/ability to join a study might be overcome amongst cases because they recognise the value of research into what affects them; they make a special effort that might create differential participation according to characteristics other than those upon which the study seeks to match).

3. **Confounding** – if factors other than the exposure of interest are affecting the outcome under study, one must question how this influences interpretation of the exposure-outcome relationship and account for this appropriately.

Confounding is a term used too freely (often without formal definition by those using it), and we will consider this issue later. Selection and confounding biases are often conflated, especially in methods employed within epidemiology that seek to overcome bias in observational analyses. Although epidemiological study designs are well-established in exercising a 'top-down' approach

of deliberation and care in minimising biases (mainly selection and confounding biases) – achieved by controlling data generation to emulate the principles of an RCT – there remains the challenge of study conduct and data analysis that yield imperfections. This is illustrated by an article in *Significance* (the official 'magazine' of the UK Royal Statistics Society): of 52 epidemiological claims that were subjected to an RCT, none were replicated; and in 5 of the RCTs the exposure effects were actually reversed[3]. From such stark evidence, it would be reasonable to conclude that the scientific methods adopted by epidemiology are not well executed or simply don't work!

Summary

In practice, the limitations and challenges of epidemiological study conduct and data analysis are well understood, and there are statistical methods available that have sought to address specific concerns about biases that persist despite idealised study designs. For instance, generalised linear regression models assume error only within the outcome, whilst most covariates in epidemiology will likely possess errors of some form. Ignoring this creates bias[4], and this is why the error-in-variables method has been proposed[5]. Propensity scores[6] are also adopted to address selection bias and confounder bias, though this method conflates the two issues. Alternatively, instrumental variables[7] are an improved approach to address potential confounding within a more formal causal framework[8].

Graphical model theory underpins a formal causal framework but has only become established as robust in the last couple of decades, and remains a work in progress. At present, the use of Directed Acyclic Graphs (DAGs; much more on these later) provides considerable clarity in revealing what we understand by 'statistical adjustment' in the context of causal inference; increasingly, we see how wrong we have been in the past in what we have hitherto thought to be perfectly acceptable. The next few lectures illustrate this.

2. CAUSAL INFERENCE & DIRECTED ACYCLIC GRAPHS

Learning objectives

- Understand the concepts of a directed acyclic graph (DAG)
- Know the correct definition of confounding within a causal inference framework
- Define the different roles that covariates can play in a multivariable regression model

Causal Path Diagrams & Directed Acyclic Graphs (DAGs)

Causal path diagrams are a visual summary of causal links amongst variables based on *a priori* knowledge, understanding, and – in the case of the relationships being tested in an analysis – conjecture. This visual summary of variable interrelationships is used in causal analysis and developed for use in expert-systems research[9]. Such diagrams are increasingly being adopted in the epidemiological community[10], yet they remain relatively novel and considerably underutilised.

Causal path diagrams may be used in a variety of ways: to think clearly about how the exposure, outcome, and potential confounding variables are causally related; to communicate these causal inter-relationships to the reader; to identify, thereby, which variables are important to measure; and to inform the statistical modelling process – particularly in the identification of confounders, mediators, and competing exposures (three roles considered in greater detail later). Causal path diagrams are the basis of a formal theoretical framework in which causal relationships can be identified and evaluated. The simplest kind of a causal path diagram is a **directed acyclic graph** (DAG).

DAGs consist of 'nodes' (or 'vertices') that represent variables (e.g. X, Y) and 'directed arcs' (or 'directed edges') in the form of arrows that depict direct causal effects (e.g. X→Y). To describe relationships between variables in such a diagram, we often read them like an ancestry tree and use kinship terminology. For example, in the diagram X→M→Y, M is a *child* of X and X is a *parent* of M; M and Y are *descendants* of X, and X and M are *ancestors* of Y. Importantly, a causal path diagram is only called a **directed** acyclic graph (DAG) if no variable is an ancestor of itself (i.e. no loops exist). Arrows in a DAG reflect *a priori* assumptions about cause and effect within the specific context concerned, some based on firm knowledge/understanding of actual (or likely) relationships between variables, others based on robust empirical evidence (preferably from a source external to the dataset under examination; see note, below), and others on entirely speculative hypotheses (including the specific relationships being examined in the analyses).

Note: These assumptions cannot (and therefore must not) be inferred empirically from the data on which the analyses are to be conducted, but are required *a priori* to select and interpret the correct statistical model.

Despite the potential visual complexity of some DAGs (particularly those with more than a handful of variables/nodes), they are nevertheless an oversimplification of the causal relationships amongst variables. A DAG does not, for example, indicate whether: an effect is harmful or protective; **effect modification** (otherwise recognised as statistical interaction, which we cover in detail later) is occurring or not;[11] or a cause is sufficient or necessary[12]. DAGs correspond to a network of variables with probability distributions (realised as the covariance structure amongst all variables).

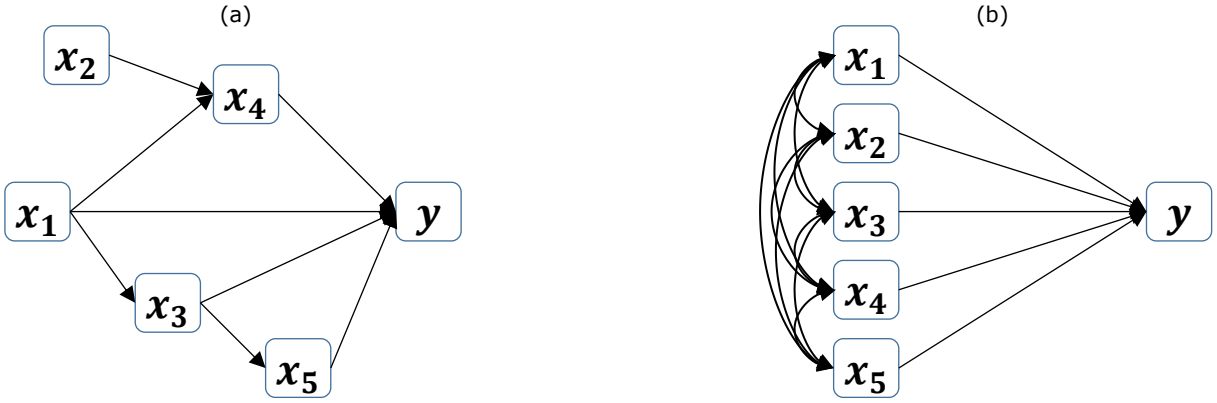
Note: There is not an exact 1:1 correspondence between a DAG and a dataset, as there are always multiple network probability distributions/covariance matrices that fit a DAG, and there can be multiple DAGs that correspond to a network probability distribution/covariance matrix. We should be mindful of the distinction between the *nonparametric* representation of postulated/hypothesised causal relationships, as captured by a DAG, and the *parametric* realisation of the variables and their relationship as described by a probability density function (PDF)/covariance structure.

One of the key strengths of a DAG is that it enables researchers to think clearly and logically about their research question(s), and to make explicit their assumptions about the relationships between each pair of variables. This visual summary (whether in the form of a DAG diagram, or an alternative representation of a DAG) can then be used to communicate these inter-relationships to other researchers, and hence it is easy to identify if, for example, any potentially important variables are missing from the DAG or whether any of the relationships suggested therein are controversial or contentious.

Limitations of the linear model

It is important to recognise that a DAG represents what is perceived to be happening causally (i.e. a hypothesised sequence of events that might fit observed data), yet when data are examined in a linear regression model (which is not path modelling) the implicit variable interrelationships are not constrained. This is illustrated in Figure 1. The *a priori* perception of causal relationships postulated by the researcher are depicted in the DAG in (a), which looks very different to how the variables are 'perceived' by a linear model, as shown in (b).

Figure 1: The relationship between outcome y and five covariates $x_1 \ldots x_5$: (a) their hypothesised causal relationship within a DAG; and (b) their covariance relationship in a linear model



In Figure 1, the relationship between outcome y and five covariates $x_1 \ldots x_5$ are considered. In (a), directed arrows signify the presumed causal links and the **absence of arrows** depict explicit assumptions of **no direct causal relationship** (in either direction). In (b), directed arrows not only link all covariates to the outcome, but bi-directed arrows (depicting correlation, i.e. direction of any cause or correlation unknown) also link all five covariates to each other. The linear model represents a one-to-many relationship without constraint, and the covariance matrix amongst the five covariates and the outcome (denoted $\boldsymbol{\Sigma}$) is freely estimated.

It is helpful to think about time (flowing from left to right) in line with causality, but this must be derived using *a priori* knowledge that is (as we've stressed earlier) *external* to the dataset. Using graphical model theory, if the data are consistent with the DAG, causality can be inferred and the extent of causal effect estimated. While causality cannot unequivocally be proven in observational

studies, even when using longitudinal data, it is nonetheless plausible to assume *potential* causality and use graphical model theory to evaluate this wherever possible.

DAG development

A careful and principled approach to developing a DAG should always be adopted. We wish to avoid a situation where one has only a vague idea of the potential causal structure, investigates the data via bivariate correlations or linear modelling, and then uses any discrepancies between the data and the DAG to revise the DAG in a cavalier fashion (i.e. with no regard for known or even likely causal relationships). Where study data are at odds with relationships deemed substantive and highly plausible, one should not immediately revise the DAG but instead seek to explain the discrepancy. Less certain relationships affirmed by the data are a sign of being on the right track, though it cannot provide guarantees that the DAG is correct. Where less certain relationships are not affirmed by the data, such relationships might not be substantive and the DAG may be valid but there are no guarantees. Relationships weakly speculated might be discarded, though this introduces strong assumptions and in general arcs are to be favoured in the presence of doubt. There is theoretical reasoning behind this strategy.

Drawing DAGs is not straightforward. Given a handful of variables and 20 researchers, each would likely come up with a different DAG if unaccustomed to drawing DAGs and left to their own devices – we'll have a chance to test this assertion in the workshop! We may construct a DAG using the online tool DAGitty (http://www.dagitty.net/) or the **R** package `dagitty`[13]. It takes considerable practice and careful thought surrounding the meaning of variables considered within the DAG before a consensus might be achieved amongst researchers. Once a DAG is obtained, its key role is the determination of what is confounding.

The role of covariates in multivariable regression

We have identified three broad areas of potential bias in causal inference in epidemiology: error in covariates, differential selection, and confounding. All three biases are important and are affected by context, but if one seeks causal inference, confounding is a critical concept to understand.

Confounding

Confounding may exist at many levels (e.g. the individual, study sample or centre, population, etc.), and most studies will identify more than one confounder. There are theoretical mechanisms by which a **sufficient set of confounders** is derived using DAGs (incorporated in the online tool DAGitty (http://www.dagitty.net/) and the **R** package `dagitty`[13]) to determine the correct causal inference of an exposure. One should always interpret a statistical regression model in conjunction with the DAG that determines which covariates are to be included in that model. If the DAG is erroneously determined (or worse, specified *post hoc* based on the statistical model), interpretation of the model is likely to be erroneous. Generally, despite its challenges, confounding can be addressed robustly if the correct DAGs are devised, though this must be done *a priori* to undertaking any modelling and executed systematically within a causal framework – a framework in which confounding has a very precise and specific definition.

There have been various attempts at defining confounding, which broadly divide into two camps: 'comparability-based' and 'collapsibility-based'[14]. In terms of the former, confounding is said to occur when there are differences in the risk of the outcome (i.e. the disease and/or healthcare

practice) in the unexposed and exposed populations that are not due to the exposure, but due to non-exposure variables that may be referred to as confounders. This results in bias in the estimate of the effect of a particular exposure on the outcome[15]. The second definition is founded on the premise that in the analysis phase of a study, confounding may be reduced or eliminated by adjusting the analysis for, or stratifying the analysis by, potential confounders[15]. The latter definition is based solely on statistical considerations, and confounding is said to occur if there is a difference in the unadjusted and 'collapsed' estimates of the effect of exposure on disease; estimates are said to have been adjusted for or stratified by the potential confounder. Although both camps are sometimes considered indistinguishable, if confounding is correctly considered to be a causal concept, rather than a statistical concept, the comparability-based definition is to be adopted[16].

Based on graphical model theory, and consistent with the 'comparability-based' definition, the accepted view for a variable to be a confounder within a causal inference framework is that it must be[15;17]:

- a cause of the outcome in unexposed people;
- a cause of the exposure; and
- unaffected by the exposure (i.e. not on the causal path from exposure to the outcome; covariates operating in this fashion being termed 'mediators').

DAGs are invaluable for identifying variables as genuine confounders. In DAGs, we can easily recognise confounders as those variables that are ancestors of both the exposure (X) and outcome (Y) via two independent paths; for instance, in $X \leftarrow C \rightarrow Y$, the variable C is a confounder but in $C \rightarrow X \rightarrow Y$, it is not. This is the strict definition of confounding, though careless use of the term 'confounder' is often adopted to describe what we now define as '**competing exposures**' and '**mediators**', which we cover next.

Note: In a linear model, confounders are correlated (i.e. *collinear*) with the exposure, which is why adjustment for confounders modifies the estimated exposure-outcome association. This is an example of a situation in which collinearity is a *good* (or at least, a *useful*) thing.

It is not always possible or necessary to measure and adjust for all known confounders. DAG graphical model theory can be applied to search for covariate sets that qualify as 'adjustment sets' that remove all confounding. The graphical rule used to find such sets is known as the 'back-door criterion'[18] and is implemented automatically in the online tool DAGitty (http://www.dagitty.net/) and the **R** package `dagitty`[13].

Epidemiological criteria used to check if a variable is classified as a confounder should be based on the comparability definition (as described earlier), though this restricts how variables in a multivariable linear model might be viewed if causality were to be inferred. The more liberal use of 'confounder' is therefore not permitted, and new terminology is needed for the more narrowly defined role of different variables in multivariable linear regression. We introduce more definitions for these variables, and for each we describe a DAG for illustration.

Proxies

We first introduce the concept of a 'proxy' variable – one that is recorded and may act on behalf of another variable, which may be recordable but not present in the study data (e.g. because it

was overlooked during the initial study design) or not recordable and therefore 'unobserved' (or 'latent'). Proxies are important as they enable us to internalise a correspondence between what we have in our data and what we seek to frame in terms of 'real-world' concepts. For instance, 'education' may be reflected differently – e.g. highest educational attainment or length of time spent studying fulltime or part-time, whichever is available in our dataset – yet such measurable variables are only proxies for 'education' that enable researchers to describe relationships between what 'education' encompasses conceptually in terms of cause-and-effect with respect to other variables in their data. There is unlikely to be a perfect correspondence (perfect correlation, statistically speaking) between highest level of education attained and length of time spent studying fulltime or part-time. Nevertheless, both measures, despite being imperfect (in effect suffering 'measurement error') allow us to capture the essence of concepts we wish to describe to investigate potential causal relationships in our data. This may seem obvious for variables such as education, but consider age: Do we mean 'chronological age' or 'biological age'[19]? Furthermore, educational attainment in early life may determine biological age later in life[20]. Considering a DAG for these variables, is it obvious which is directly observed and which is a proxy?

Less obvious, though equally important, is that some variables in a causal chain may be missing (i.e. not recorded in our dataset), yet their implicit presence is central to the correct drawing of causal paths linking variables. Consider the three variables for an individual for whom we have information regarding their parents, their diet during childhood, and their BMI when entering adulthood: 'parental education' (PE), 'childhood diet' (CD), and 'adult obesity' (AO). It seems reasonable to draw a DAG as PE→CD→AO, where we surmise a causal chain from PE through CD to AO, since more educated parents are more likely to provide the kind of childhood environment, including dietary influences, that lead to a lower risk of obesity as individuals enter adulthood. Following this logic, if the information regarding individuals' diets were absent from the study dataset, we may surmise PE→AO. This does not mean that parental education *directly* causes obesity in offspring's adulthood, but that the proxy of 'childhood diet' is not present, so we drop CD from the DAG and retain a causal arc from PE to AO, since CD is a descendent of PE and an ancestor to AO. Many hypothetical proxies may exist as decedents of one variable and ancestors to another variable, thereby linking the two by proxy.

It is important to recognise how variables in our data may subliminally supplant a more complex array of factors we are interested in whether clinically, biologically, or from an ecological perspective – and necessarily so, as it facilitates the exposition of what are typically complex research questions. The implicit distillation processes we go through to arrive at the models we employ in addressing our research questions are as limited (i.e. 'approximate') as the models themselves in representing 'truth' (or, in other words: "all models are wrong, but some are more useful than others"[21]). We are prone to overlooking these simplifications of what our data represent and worrying only of completeness, measurement error, and robustness of the statistical methods used in their analysis (though the latter, too, is also often overlooked).

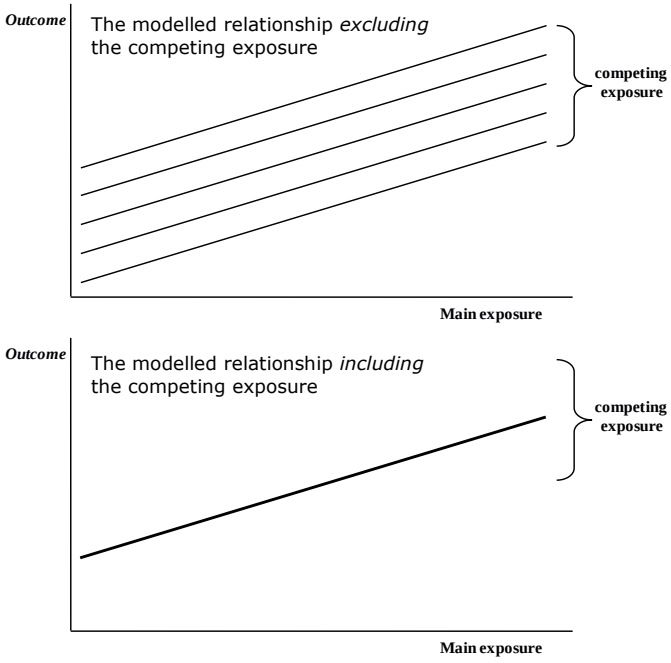
Competing exposure

A competing exposure is strictly not a confounder, though researchers often conflate confounders with competing exposures. For a variable to be considered a **competing exposure**, it must be:

- a cause of the outcome, or a proxy of a cause of the outcome (i.e. an ancestor to the outcome);

- ***not*** a cause (or a proxy of a cause) of the main exposure; and
- unaffected by the main exposure (i.e. not a descendant thereof).

With no other variables in the linear model, the estimated association (slope) between an exposure and an outcome is unaffected when the linear model includes a competing exposure, since orthogonal covariates do not impact each other's estimated coefficients.



Competing exposures are unlikely to be completely orthogonal to the main exposure, especially when they are continuous variables, so their inclusion will change the estimates slightly.

Always, however, **precision is improved** because some of the outcome uncertainty is effectively 'explained' by the competing exposure.

Note: Within the population, the competing exposure and the (main) exposure are assumed to be **causally unrelated** but may nevertheless be **correlated**. A competing exposure may be correlated with the (main) exposure in the **study sample** for one of two reasons:

- Although the main and competing exposures are **causally unrelated** at the population level, in the study data they may exhibit a non-zero correlation due to chance sampling. Were the study repeated several times, *on average* the estimated association between the main exposure and outcome is correct (hence there is no statistical bias), but for any one study sample with chance correlation, the estimate will be modified away from true; or
- There is an ancestor (observed or unobserved) that causes both, creating a **correlation** at the population level and, therefore, a likely correlation within any subsample. Inclusion of the competing exposure will modify the main exposure-outcome relationship, which is desirable, as the competing exposure is then actually better understood as a **proxy confounder** (see next section).

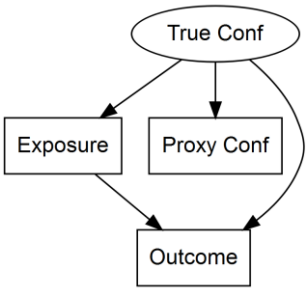
If a competing exposure is correlated with the main exposure in a study sample and this is due to chance (i.e. the first instance above), the competing exposure should not be included in the model, as inappropriate modification of the main exposure-outcome estimate occurs. This would trump any advantage of improved precision because, although the estimate would be more precise, it would be incorrect!

If, however, a competing exposure is correlated with the main exposure due to an ancestor variable causing both exposures (i.e. the second instance above), inclusion of the competing exposure remains favourable (to adjust for proxy confounding), thereby removing bias and improving accuracy, whilst improved precision will also result (as a competing exposure) – a win-win!

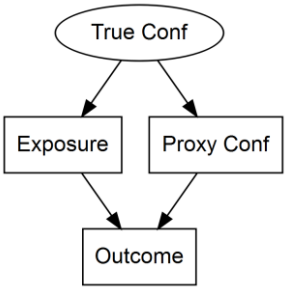
The only way to be sure that inclusion is favourable from a confounder adjustment perspective is to use a DAG for all variables considered relevant, available or unobserved, and determine all

possible adjustment sets (which can be done automatically thanks to the online tool DAGitty [http://www.dagitty.net/] or the **R** package `dagitty`[13]). The decision as to whether to include the competing exposure in the linear model is determined by: examination of the *a priori*, appropriately determined, DAG (which is non-parametric and can therefore only indicate which adjustment sets are appropriate rather than the exact nature of the variable relationships); in conjunction with knowledge of the data and its parametric sampling properties (which may indicate a non-zero correlation arising in the sample even when the DAG indicates the correlation should be zero in the population).

Proxy confounder

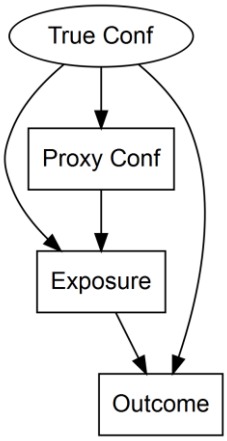


In the situation where the (main) exposure and any competing exposures are not directly causally related but are correlated due to a common ancestor variable that is causally related to both (maybe unobserved; see right), the competing exposure is a '*proxy* (of the ancestor that is a true) *confounder*', i.e. a **proxy confounder**.



Note: The true confounder need not be a parent of the outcome. This might seem to contradict the definition of a 'true' confounder, but were we to remove the proxy confounder from the DAG, the link between the true confounder and the outcome becomes direct; being an ancestor is sufficient to be a cause of the outcome.

Proxy confounders are not themselves confounders, but lie on the causal path between confounders and either the (main) exposure or the outcome (but not both; else it would be a confounder).



Proxy confounders are useful where true confounders are not observed and the closest we have to assessing the impact of unobserved true confounders is through their influence via their proxies.

Mediators (on the causal path)

A common challenge in epidemiology involves the (in)appropriate treatment of mediators in linear regression models. As mentioned earlier, mediators are variables that lie on the causal path between the exposure and outcome.

Including mediators in a regression model can yield a statistical artefact sometimes known as **Simpson's Paradox**, where it typically arose from the assessment of categorical covariates and did not need to invoke linear modelling. We later introduce **Lord's Paradox**, which was originally illustrated amongst both categorical and continuous covariates, though can arise with any combination of covariate types. We also cover **suppression**, which is typically described amongst continuous covariates but may arise with categorical data. All three paradoxes became known separately in different contexts but are manifestations of the same phenomenon termed the **reversal paradox**.[22] Henceforth we will describe all these forms of 'paradox' as the 'reversal paradox'. The phenomenon was so named because the adjustment for a mediator *can* (though does not necessarily) give rise to a sign change in the exposure model coefficient. More importantly, even when no sign change occurs, there can be substantial bias in the inferred model coefficient estimate[23].

The reversal paradox may arise in linear modelling for any combination of categorical and continuous variables. The problem of statistical adjustment for mediators is ubiquitous yet controversial and highly contested. There is, however, no actual 'paradox', only limited comprehension of the causal framework in which a linear model is to be interpreted[24]. Despite a plethora of articles outlining how it is challenging to interpret findings of regression models with respect to causal inference if mediators are included[22;23;25-28], the practice of including them inappropriately in linear models persists. This is in part because some find its implications overstated[29], though likely because many do not fully grasp causal theory[24;30], which explains the paradox and helps interpret what is happening.

Note: Estimated coefficients in a linear model that includes one or more mediators are not **statistically biased**, as they are correctly estimated – the mathematics of the linear model are expedited robustly. Instead, what is attributed to model coefficients is **inferential bias** (i.e. an incorrect inference made about the causal relationship between an exposure and the outcome).

Summary

Inferential bias occurs whenever insufficient care is taken to build a linear model set within a causal framework, where it is critical to select an appropriate combination of covariates to optimally 'adjust' for confounding. If the adjustment set of model covariates for a specific exposure-outcome relationship is not appropriate (as per the idealised DAG for that context and associated datasets) no meaningful causal inference can be made of the exposure-outcome model coefficient.

This leads nicely onto the challenges of selecting appropriate covariate subsets and interpreting robustly the coefficients of a multivariable linear model from a causal inference perspective.

3. DRAWING DAGS

Learning objectives

- Rehearse the principal terms used to describe each component of a DAG
- Learn how to systematically approach the drawing/specification of a DAG

Definition and terminology

Within causal inference, DAGs are graphical, nonparametric representations of hypothesised causal relationships between measured ('observed') and unmeasured ('unobserved' or 'latent') variables. Current convention is to represent measured variables as squares or rectangles and unmeasured variables as circles or ellipses, although this is not universally applied. These representations of variables are termed 'nodes' (or 'vertices'), and the causal paths between variables are represented by unidirectional arrows termed 'directed arcs' (or 'directed edges').

Three key characteristics of DAGs are that:

- causal paths between variables **must be** unidirectional (i.e. each of the variables connected by a causal path can only operate as either cause or effect, and not both);
- a variable **must not** cause either itself or one of its own causes (i.e. there should be no cyclical paths, hence the name 'directed **acyclic** graph'); and
- while a direct path between two variables only indicates the **possibility** that these variables are causally related (even if only to a modest extent); the absence of a direct path between two variables reflects the absence of any such causal relationship (i.e. greater **certainty** and importance is afforded the absence of a causal path than the presence of one).

Epidemiological utility – past, present and future

We have seen how DAGs have substantial utility for displaying – and supporting robust analyses of – hypothesised causal relationships. DAGs facilitate what might be termed a 'causal gaze' – a perspective from which complex (causal) processes can be simplified, characterised in graphical form and then examined, disentangled, debated and resolved using an established framework of rules (including the three key characteristics listed above).

DAGs also facilitate the identification of variables operating in very specific ways within any hypothesised causal system, each of which requires particular attention when designing statistical models to generate causal inference. As described in the preceding session, these include:

- the specified (or 'main') '**exposure**' (the putative cause within the '**focal relationship**' under examination);
- the specified '**outcome**' (the putative effect / consequence within the '**focal relationship**' under examination)
- potential '**confounders**' (covariates relating to events, processes, characteristics which, as specified, occur **before** *both* the exposure *and* the outcome, and are therefore **potential** causes of both);
- likely '**mediators**' (covariates relating to events, processes or characteristics which, as specified, occur **after** the specified exposure but **before** the specified outcome, and are therefore **potential** consequences of the specified exposure and **potential** causes of the specified outcome); and

- '**competing exposures**' (covariates that are causally **unrelated** to the specified exposure but which precede, and are therefore **potential** causes of, the specified outcome).

By identifying variables operating in these ways within the hypothesised causal system, DAGs have extensive utility in statistical modelling for causal inference by ensuring that models:

- identify, and adjust for, those covariates specified as potential confounders;
- do **not** adjust for covariates specified as likely mediators (since the adjustment for such variables can create bias due to the 'reversal paradox'[23]); and
- can identify and adjust for covariates specified as competing exposures, wherever such adjustment strengthens the model produced.

The application of '**graphical model theory**' to DAGs[31] can further enhance adjustment for confounding by identifying any alternative '**minimally sufficient adjustment sets**' of covariates specified as potential confounders[9]. This can be of great practical value in those circumstances where: not all of the specified potential confounders have been measured; or not all of the specified potential confounders can be measured with reasonable accuracy and precision (or within the resources available).

Beyond these 'early benefits' of DAGs (i.e. improving the transparency of *a priori* hypotheses; reducing inappropriate adjustment for mediators; and enhancing the selection of confounders for adjustment), DAGs also have substantial potential utility for: identifying and estimating the extent of unobserved confounding (where the DAGs involved permit this); evaluating whether any given DAG (as specified) is consistent with the observed dataset(s) it was intended to represent[13]; and elucidating invalid or inappropriate analyses.

Conceptualising variables and contextualising cause

Although DAGs can sometimes offer simple representations of what might otherwise be complex causal processes, many can be challenging to draw (or, rather, to 'specify'), not least when:

- the variables involved represent poorly defined and/or understood concepts/constructs;
- the variables, though measured at one point in time, reflect events, processes or characteristics that occurred at previous points in time; and
- the causal processes the DAG is intended to reflect are influenced by the context(s) in which these occur.

Hypothesising the potential causal relationships between each of the constituent variables (be they manifest or latent) requires that we not only recognise precisely what each variable represents (be that an event, a process, or a characteristic), but also that we have substantial understanding of each potential causal relationship based upon clear theoretical principles and/or robust, *external* empirical evidence. This can be extremely challenging, especially in hypothesised causal systems where there is incomplete understanding, limited robust *external* empirical evidence, or where the theoretical principles involved are unclear, uncertain or contested. Nonetheless, even under these circumstances, 'temporality' (i.e. the simple rule that the past precedes the present) can often provide a sufficient theoretical basis upon which DAG specification can proceed, providing it is possible to identify the temporal sequence of the variables involved. Thereafter, there is no reason why alternative DAGs (particularly, and preferably, when specified *a priori*) might be specified that

reflect specific uncertainties and therefore guide causal inference analyses using DAG-informed sensitivity analyses.

Determining the temporal sequence of variables within a DAG requires establishing the temporal relationship of measurements operationalised as nodes that are fixed in time either: (a) by nature of the variable concerned (i.e. where the variable is '**time-invariant**' and varies only across subjects/participants and not over time; e.g. sex or place of birth); or (b) by the specific point in time at which the variable concerned was measured (i.e. where the variable is '**time-variant**' and varies not only across subjects/participants but also over time; e.g. body mass or food intake). Importantly, every measurement of a time-variant variable captures not only the value prevailing at the point of measurement, but also the cumulative 'experience' of that variable over the time preceding measurement (such that the measurement made might be considered to represent a value that has 'crystallised' at, or up until, that point in time).

The precise time at which a time-variant variable (and the concept/construct this represents) is 'crystallised' is crucial for considering where it should be placed in the temporal sequence of nodes that form a causal DAG. This is because temporality is key to establishing which variables (as manifestations in time of the 'crystallised' properties they reflect) can plausibly act as **potential** causes of other variables (given that only past nodes can cause subsequent nodes). Indeed, the very notion of time-variant variables – which may reflect properties from **either the present or the past (or both)**, that have accumulated over time – can make them especially difficult to position within a DAG (both conceptually and functionally). A simple example of such a variable might be body height which might be considered a time-variant variable when measured during childhood, but which might appear time-invariant when measured in adulthood (having crystallised at the end of adolescence, thereafter remaining the same until the decline in height commonly accompanying senescence later in life).

The causal relationships between variables (whether time-variant or time-invariant) may also change between contexts, such that a valid causal relationship in one context may be reversed in a second, or entirely impossible/implausible (and therefore absent) in a third. Drawing DAGs therefore requires not only careful thinking about the meaning of all of the constituent variables, but also how these are likely to be ordered, in time, within the specific context being modelled – a context that extends not only to the specific historical, social and physical environment concerned, but also to the very different 'analytical contexts' that exist for different study designs, sampling strategies, and data acquisition processes.

Understanding any given variable, what this purportedly measures, and what this means in any given context, is therefore both challenging *and* critical to correctly specifying DAGs that are capable of informing robust statistical models of hypothesised causal relationships. There may be instances in which the level of ambiguity or a lack of knowledge and understanding means there is little confidence to support accurate specification of even the most hypothetical DAG. Yet the impossibility of knowing, *a priori*, everything necessary about the processes involved in any causal system does not mean that the resultant DAG (specified *in the absence of definitive evidence*) has nothing to offer to strengthen our confidence in causal inference modelling. This is because, while challenging to specify and impossible to perfect, DAGs nonetheless make the process of causal estimation far more transparent to both the analysts concerned and to others. By helping analysts

to identify (and hence avoid) some of the more obvious (and sometimes less obvious) errors that influence the analysis of observational data for causal inference, even 'uncertain' DAGs can help improve the analysis of causal inference.

Drawing DAGs in four simple steps using temporal logic

Notwithstanding the conceptual and operational issues considered above, there are four simple rules (based on the unassailable 'temporal logic' that the past precedes the present) that can help to improve the drawing/specification of DAGs to represent hypothesised causal processes.

- *All nodes should be considered as potentially 'time-variant' measures of the variable they represent:* this ensures that the properties attributed to **measured** variables include those that may have crystallised prior to the time at which the variable was measured.
- *Simultaneously crystallising variables are likely to share common (latent or manifest) causes:* this allows for any such 'contemporaneously crystallising' nodes to be **correlated** without being specific about a direct causal link, nor having to specify the direction of any such cause (if present).
- *Only preceding nodes act as causes of subsequent nodes:* this requires nodes acting as causes to have properties that crystallised **before** those of any nodes they affect.
- *Temporality confers the **potential** for causality:* this means that causal paths (i.e. arcs or edges) should only be missing within a DAG where these: do not follow temporal logic; or where there is robust, *external* empirical evidence that the given causal path does not exist.

These four rules can be translated into a series of tasks that greatly facilitate the specification of DAGs based on all constituent variables (whether observed or unobserved) that are thought to be relevant to the focal relationship under examination:

- **First**, determine **when** each observed variable (regardless of when measured) was likely to have 'crystallised'; then specify **when** each unobserved variable is considered (theoretically) to have crystallised; and arrange both sets of (observed and unobserved) variables in a temporal sequence, allowing for groups of variables that crystallised at the **same** point in time to be situated contemporaneously;
- **Second**, for each group of contemporaneously crystallised/situated variables, add a new latent (i.e. unobserved) variable operating as a common cause temporally situated **immediately** preceding the contemporaneously crystallised / situated group of variables.
- **Third**, add directed arrows from **all** preceding variables to **any** subsequent variable(s), ensuring there are no missing arrows from any preceding variable to any subsequent variable.

The first three steps generate what is termed a '**forwardly saturated DAG**' (meaning that it includes all possible causal paths between preceding and subsequent variables). When drawn in a straight line (e.g. from left to right, from past to present), with variables arranged in the order in which these crystallised, and with causal paths delineated using curved lines, such DAGs often take on the appearance of an 'onion' (hence the colloquial term '**onion DAG**')[32].

Importantly, a **fourth step** may be required when there is sufficient evidence to warrant excluding a directed arrow between a preceding and subsequent variable, thus:

- **Fourth**, remove **only** those directed arrows between variables where these do not follow temporal logic (this should not occur if the third step, above, has been correctly implemented)

or where there is sound knowledge or robust, *external* empirical evidence that the given causal path does **not** exist.

Summary

After highlighting the early benefits of DAGs (in facilitating the conceptualisation of causal systems and processes and helping to reduce a range of common flaws and errors in the modelling of causal systems), we also examine several implicit and explicit conceptual and contextual challenges to drawing (or 'specifying') DAGs. These challenges relate to both: the causal meaning of what constitutes a 'variable' (and the 'nodes' used to represent these as markers of past or present events, processes, or characteristics); and the important role that context plays in determining what variables mean, and how they are conceptualised and operationally specified. The four key rules outlined, based on temporal logic, can be applied using four simple steps to draw/specify DAGs consistently, thereby improving intra- and inter-analyst reliability and reducing the potential for error.

4. STATISTICAL ADJUSTMENT IN MULTIVARIABLE LINEAR MODELS

Learning objectives

- Know how DAGs inform covariate selection in a multivariable regression model
- Know when to adjust / not to adjust for mediators in a multivariable regression model
- Learn how to interpret correctly the coefficients of a multivariable regression model

Causal interpretation of multivariable linear models

In causal inference, two variables are special:

- **exposure** (or treatment); and
- **outcome** (or endpoint).

All other variables are **covariates**. As we have seen, covariates have a variety of different roles from a causal inference perspective: they can be **mediators**, **confounders**, **proxy confounders**, or **competing exposures**. If a suitable subset of covariates can be identified that removes confounding, we may proceed to estimate our causal effect using a multivariable linear model.

In regression models, there are only two types of variables:

- dependent variable (DV) and
- **independent variables** (IVs, predictors, or covariates).

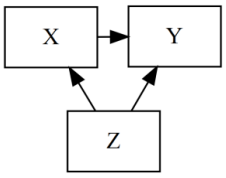
No further distinction is made between the IVs – specifically, the exposure is by no means a "special" IV and is treated just like any other covariate. Thus, there is a conceptual mismatch between causal graphical model theory (as depicted by DAGs, which lead us to formulate a multivariable linear model that highlights the exposure-outcome relationship adjusted for confounding) and the standard perception of a regression model. This conceptual mismatch often leads to misinterpretation of the results from a multivariable linear model.

Table 2 Fallacy

One particularly widespread misconception is known as **mutual adjustment**, recently called the 'Table 2 fallacy'[33], since the first table in most epidemiological articles usually describes the study data and the second table reports the results of a multivariable regression model where the erroneous efforts to illustrate mutual adjustment often appear.

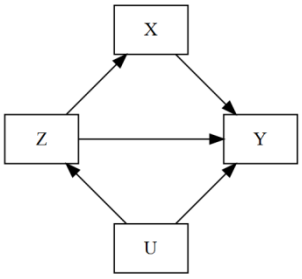
To illustrate the fallacy, let us assume that we wish to estimate the effect of X on Y. We know (e.g. from a DAG) that there is only one confounder, Z, so we run the regression Y~X+Z. If our background knowledge and the statistical assumptions of the regression (e.g. normality) hold, then the coefficient of X estimates the **total causal effect** of X on Y. The 'Table 2 fallacy' is the belief that we can also interpret the coefficient of Z as the effect of Z on Y; indeed, in larger models, the fallacy is the belief that all coefficients have a similar interpretation with respect to Y.

To see why this is not true, look at the DAG that matches our scenario: Z→X→Y & Z→Y (see right). With respect to the X→Y effect, adjustment for Z removes all confounding, but what does including X in the model mean for the effect of Z on Y?



As we can see, X is a mediator of the Z→Y effect, but adjustment for a mediator is erroneous when estimating the total causal effect; the Z coefficient in our model cannot be interpreted as such. Instead, we could interpret it as the 'direct effect' of Z on Y when X is held constant, and this may be stronger than, weaker than, or opposite to the total effect. It would seem, from this example, that we can at least interpret every coefficient as a causal effect: some total and some direct.

To see that this can also fail, let us add another variable to our DAG. We include U, which affects both Z and Y (see right). Despite the addition of this new variable, it is still sufficient to adjust for Z to unconfound the X→Y effect, so the validity of the X coefficient is unchanged – can you see why? Upon examining Z in this situation, however, we encounter difficulties.



The new variable U acts as a confounder of the Z→Y relationship, which means that we would have to interpret the Z coefficient as a 'direct effect that is *confounded* by U' – not exactly a helpful interpretation. Indeed, no single multivariable linear model could ever estimate the causal effects of X and Z at the same time: estimating the X effect means we *must* include X in the model, but to estimate the Z effect we *must not* include X.

In general, it is impossible to identify multiple causal effects using a single linear model, and we can usually interpret at most one coefficient in such a model as a **total causal effect**. If we are interested in multiple causal effects, we need multiple (**separate**) regression models.

In the 2nd DAG, we can obtain the effect of X from the model Y~X+Z because adjustment for Z unconfounds the X→Y effect, and we can obtain the effect of Z from the model Y~Z+U because adjustment for U unconfounds the Z→Y effect. The concept of 'mutual adjustment', as often encountered in the literature, is seriously misleading and erroneous.

Statistical adjustment

Within observational research, it is important to adjust for confounding to reduce potential biases. Other forms of adjustment may be undertaken, e.g. for competing exposures, which are not true confounders but can improve model precision (recall: some competing exposures might also double as proxy confounders). Adjusting for mediators (variables that lie on the causal path from exposure to outcome) presents a challenge, as this may bias the intended model *inference*.

We now use DAGs to examine carefully when and how to make 'appropriate' statistical adjustment for mediators in a linear regression model. To do this we must recognise three key ingredients to the application and interpretation of multivariable regression models:

- ***causality*** – the framework in which confounding is defined;
- ***intervention*** – whether real or hypothetical, as a basis of thinking about what has meaning in relation to the research question that drives interpretation of the model coefficients; and
- ***context*** – a 'catch-all' for remaining issues, but important for the recognition of extraneous factors that validate or challenge the appropriateness of methodologies adopted; an example is how we understand the abstract meaning of variables in our DAG (discussed at length in our first example that follows).

In 1995, Judea Pearl formulated a new calculus for application to causal graph theory coined *do*-calculus[34]. Pearl's calculus facilitates identification of causal effects in non-parametric models as well as proving useful in mediation analysis[35], transportability[36], and the recently emergent domain of meta-synthesis (the fusing of empirical results from diverse studies conducted on heterogeneous populations, under different conditions, to synthesize an estimate of a causal relationship in some target environment). We do not consider this calculus in detail but borrow the '*do*' component, i.e. the concept of intervention. When considering the implications of causality in model selection and model interpretation, it helps to think about the role of intervention, either real or hypothetical. Drawing meaningful inference in observational research from a linear model then boils down to identifying the context in which inference has utility. This is best realised by asking: *What is the causal consequence I am interested in?* This helps target an intervention that corresponds to the research question.

To illustrate, we consider two contexts in which statistical adjustment for a mediator has a different impact (modifying the estimated exposure-outcome relationship appropriately or inappropriately) and see how this relates to a hypothetical intervention. In our first context, we consider a variable that researchers often adjust for because they view it as a confounder, though it is a mediator. We discuss such contexts in which mediator adjustment biases the intended causal inference and is therefore inappropriate. In our second context, we consider a variable that is well-understood to be a mediator, yet adjusting for it is necessary to gain correct causal inference. We explore and explain this apparent contradiction, highlighting key differences between the two scenarios in terms of hypothetical interventions, indicating when mediator adjustment is appropriate or not.

Context 1: The relation between adult blood pressure and birthweight

In considering a potential relationship between adult blood pressure (BP) and birthweight (BW), researchers have questioned the validity of any association in part due to publication bias and/or inappropriate statistical adjustment for variables on the causal path (such as adult body size)[37], as the latter gives rise to statistical artefact called 'reversal paradox'[23]. It has also been shown that simultaneous adjustment for two or more intermediate measures of body size exacerbates this artefact[27]. Nevertheless, it is suggested that some intermediate measures (e.g. adult weight, AW) are proxies for genuine confounders that are either unmeasured or, as yet, not identified (e.g. genes that simultaneously affect BW, adult body size, and adult BP)[29]. Concern with this argument is that if intermediate body size measures are a proxy for unmeasured or unknown genuine confounding, the reversal paradox does not go away; there are adverse effects of the artefact induced by the reversal paradox and genuine bias-reduction due to adjustment for proxy confounders[25]. In many situations, it may be unclear, and even unresolvable, as to which direction and of what magnitudes these effects alter the estimated model coefficient for the main exposure; they may be synergistic (add to) or antagonistic (oppose and partly cancel out). In any event, the inferential bias from the reversal paradox never goes away.

It helps to resolve this dilemma by asking: what is the *research question*; what *consequence* are we interested in; and how might we assess this via a (hypothetical) *intervention*?

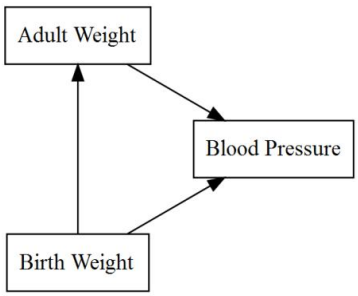
These issues are **context specific**. For instance, do we wish to understand the impact of BW *per se* or, more likely, are we interested in what BW is a proxy for? Biologically, it is unlikely that body mass at birth in a **physical sense** is at all important in relation to adult BP; rather, it is what body

mass at birth represents, what it reflects of **foetal development**, and whether this has some bearing on physiological status in later life. It is widely accepted that BW is a proxy for many things, not least **in-utero nutrition** (quality and quantity); and health of the foetus is also affected by the health of the mother (before and during pregnancy). To affect adult BP via intervention at the earliest stages of life, if BW is associated with adverse health outcomes in later life, one might seek to affect all factors reflected in the proximal value of BW. We might therefore seek to ensure mothers are fit and healthy before conception, as well as during pregnancy; we might seek to ensure mothers' diets are balanced, containing sufficient nutrients and calories for optimal foetal development; and we might seek to secure a more holistic positive environment to minimise physical and mental stress, avoid adverse lifestyle choices (e.g. alcohol, tobacco), and minimise disease exposures (e.g. measles, tuberculosis).

The complexity of BW as an exposure brings into question what it is that any unmeasured or unknown confounders confound: do they causally influence all or just some of the factors encapsulated in the proxy measure of BW? If some unmeasured or unknown confounders were genetics, for instance, how do genes determine environmental factors that influence BW? Apart from operating via biological mechanisms that drive dietary habits and/or general health-related behaviours (e.g. alcohol or tobacco addiction), many environmental influences of maternal and foetal wellbeing are determined by geographical, community, and cultural circumstances, such as the availability of foods and medicines (even in developed countries), the risk of exposure to disease or disaster – whether natural (earthquake, floods) or man-made (war) – and parochial norms in diet and lifestyle. This perhaps makes for an argument that any confounding, for which adult weight purportedly acts as a proxy, is tenuous and dilute for each potential confounder. One might argue that many other factors that may seem arbitrary, yet conveniently recordable, could similarly be considered proxy confounders and we soon become awash with possible proxies.

We might seem to overanalyse BW as an 'exposure', but this discussion serves to illustrate that the variables we use in a linear model are an abstraction of what we hope they reflect. When seeking causal inference, and thus when considering the role of various measures as confounders or proxy confounders, the **perspective** adopted is **subjective**. Most clinical variables have utility, though often only approximately encapsulating the essence of our research focus. We should remain mindful of this when undertaking linear modelling for causal inference.

Stepping back from important yet philosophical issues of context and utility / meaning of variables in our DAG, we examine what is meant by adjusting for mediators. We form a theoretically sound perspective by constructing a DAG (right), leaving aside whether intermediate body size captures confounding by proxy, and examine the BP-BW relationship as though causal, with focus on a potential intervention just before BW is measured.



We ask the question: *What is the effect of one unit change in BW on change in adult BP?* We consider this with two model scenarios: one where we have BP as the outcome and BW as the exposure variable and no other covariates (i.e. BP~BW); the other where BP is the outcome, BW is the exposure variable and we include current adult weight (AW) as a 'confounder' (i.e. BP~BW+AW, ignoring that AW is not a 'true' confounder). Critically, we assume a causal BW-AW

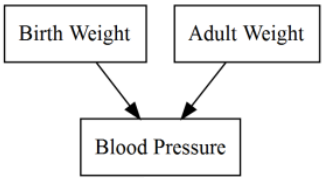
relationship (supported by the literature), else AW is not a mediator either but rather a competing exposure.

In using the model that includes AW to estimate the impact of change in BW on BP, we must evaluate the impact of change in BW on both AW and BP, along with the impact of an altered AW on BP. In using the model that includes only BW, we must evaluate the impact of BW on BP only.

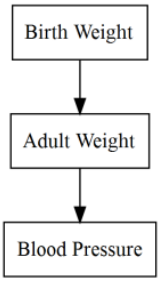
In the causal framework (BW→AW→BP and BW→BP), it is shown mathematically that the evaluated impact of one unit change in BW on BP is identical for both models yet more succinctly captured by the BW coefficient in the BP~BW model. The BW coefficient in the BP~BW+AW model does not reflect the *total* effect of BW on BP, as it must be modified by the effect of AW on BP.

From a causal inference perspective, asking: *What is the (hypothetical) intervention-generated effect of one unit change in BW on the change in BP?*, the model to yield the answer is the model with only BW included; inclusion of the intermediate AW modifies the coefficient of effect for BW away from the true *intervention-generated* effect. Since BW→BP *and* BW→AW→BP (i.e. AW is a mediator), 'adjustment' for AW in the BP~BW model alters the inference sought of the BW coefficient (which is interpreted around the idea of an intervention on BW).

Note 1: If BW is not causally related to AW, it is a competing exposure (see right) and there would be no difference between the two models in the coefficient estimated for BW and both models would capture the *total* causal effect of BW on BP correctly in the BW coefficient.



Note 2: If BW is not directly causally related to BP (see right), then adjustment for AW should completely remove the effect of BW.



Note 3: Both models (BP~BW and BP~BW+AW) are **statistically unbiased**, as they are correctly estimated; the second model suffers **causal inference bias** (the estimated impact on BP of a hypothetical intervention on BW is biased).

We thus conclude that when seeking to interpret an outcome-exposure relationship causally within a multivariable linear model, where interpretation of the exposure coefficient is predicated on an intervention at the time of (or just) before the exposure assessment, then inclusion in the linear model of mediators biases the model inference and hence its interpretation (due to the *reversal paradox*); the exposure model coefficient does not reflect the total causal impact of any hypothetical intervention on the outcome.

Context 2: Relation between sex and academic career progression

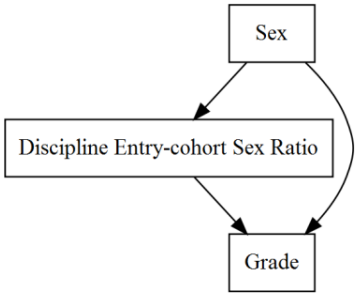
It is generally acknowledged that there are differences between the sexes, though what is due to nature or nurture is still debated[38]. It is nevertheless widely accepted in science (and increasingly accepted culturally, reflected in legislation) that, notwithstanding variation within each sex, men and women are on average no different in their potential intellectual acuity[39]. It is thus reasonable to presume that it is entirely down to cultural differences experienced throughout life that leads to sex imbalance in the pursuit of different careers. Therefore, within professions for which there is no reliance on physique, uptake of jobs and progression through the ranks should be proportionally very similar. In academia, for instance, the proportion of men and women in each discipline at each grade should be roughly equal. This is, however, far from true (across the globe in fact). In

the UK, this led to the formation of an equality charter, **Athena SWAN**: committed to advancing women's careers in science, technology, engineering, maths and medicine (STEMM) in higher education (see: http://www.ecu.ac.uk/equality-charter-marks/athena-swan/).

One metric used in raising awareness and used in monitoring the success of the Athena SWAN charter is the proportion of women at each grade, e.g. the proportion of women professors per discipline. An implication is that we can assess the 'performance' of academic institutions to 'promote sex equality' through such a metric. The academic workforce today, however, is the product of individuals' experiences over the years prior to their first appointment, including their journey through postgraduate and undergraduate training, and before that through secondary education, primary education, nursery, and home-life, along with the wider societal and cultural environment throughout their lives. When examining institutions for potential sex discrimination, we must take account of this.

Contemplating how to investigate academic institutions in their 'fairness' to promote men and women equally, we can look at the proportion of **successful appointments** by sex at each grade, and then 'adjust for' *the proportion of men and women applying each time*, though this information is unlikely to be available. Instead, we might adjust for the **proportion of men and women eligible** for each appointment by considering **discipline-specific entry-cohort sex ratios**.

One problem is that entry sex ratios may vary over time, and the lag between entry and each appointment widens with seniority of grade. For simplicity, we assume no change in discipline-specific entry-cohort sex ratios over time and consider hypothetical data for all academics in STEMM subjects comprising: academic **grade** (outcome), their **sex** (exposure), and each **discipline entry-cohort sex ratio** (mediator); see the DAG on the right.



The exposure (sex of the individual) precedes entry to any academic discipline and subsequent grades attained; each discipline entry-cohort sex ratio precedes any subsequent grade attained and lies on the exposure-outcome path. As to whether these relationships are causal must be determined. As cumulative lifecourse experiences differ by sex prior to entry into an academic career, the discipline entry-cohort sex ratio is a proxy for these experiences in the same way as birthweight was for early-life exposures. A causal link between sex and discipline, hence entry-cohort sex ratio, is therefore implicit. In the absence of any sex discrimination, discipline entry-cohort sex ratios should yield similar sex ratios in grade attainment, with proportions of each grade by each sex determined by the discipline; causality is again implicit.

In a linear model, discipline entry-cohort sex ratio is a mediator. As per the BW~BP example, adjusting for discipline entry-cohort sex ratios whilst examining the grade-sex relationship might be suspect. On the other hand, it is compelling to 'adjust' for discipline differences in the workforce sex ratios, as alluded to. To resolve this, we ask: what is the *consequence* we are interested in; and how might we assess it via (hypothetical) *intervention*? The answer to these questions helps frame the research question: *Are appointments to grade subject to sex discrimination?* The process that then takes place to address the research question occurs when the grade is attained, which is after the entry-cohort sex ratio was established. The consequence of interest is ensuring fairness in the *appointment* process. Hence, we need to adjust for the entry-cohort sex ratios because they

differ at the time each appointment is made, thereby affecting the denominator of men and women entering the selection process.

The fairness being assessed (upon which one might hypothetically intervene) occurs at the time the outcome (grade) is measured, not when the exposure (sex) is measured, and importantly *after* when the mediator (entry-cohort sex ratio) is measured. If we intervene to change establishments prone to sex discrimination, this would be to alter the appointment process, e.g. by ensuring that appointment committees are gender balanced, involving independent observers to intervene if any part of the appointment process fails to give equality to all candidates, or other such actions **at the time of appointment**. The critical point is that any intervention necessarily takes place at or just before grades are attained (or not), and therefore *after* the time when discipline entry-cohort sex ratios are established.

As per our DAG, the total causal effect of sex on appointment status in higher education comprises an indirect effect mediated by societal factors that lead to a certain entry-cohort sex ratio, and a direct effect not mediated by such factors that preceded the application. Any policy change of academic institutions in the hiring process cannot hope to change the indirect effect (e.g. gender balanced committees cannot influence the choice of toys in nursery). What is targeted is the direct effect: by adjusting for the mediator, we 'block' the indirect effect so only the direct effect remains, which is the relevant effect for our intervention question.

Summary

The key to understanding when to adjust for a mediator in a regression model is to ask when might an intervention be required that best informs our research question. If the intervention occurs after the mediator, it is appropriate to adjust. Conversely, for mediators occurring after the intervention it is inappropriate to adjust. By framing research questions in terms of an intervention, it highlights which factors confound the **intervention-outcome relationship** as opposed to the **exposure-outcome relationship**.

It is not important that factors considered for 'adjustment' are confounders or mediators if all precede the intervention point. This keeps the application and interpretation of conditional linear modelling firmly rooted in a causal framework. It is the need to arrive at causal inference that leads to such rigid ways we think about and employ multivariable linear models (DAGs aid this). We considered two contexts: one in which adjustment for the mediator was inappropriate because what was to be estimated was **total effect**; in the other context, the desired effect was the **direct effect**, and so adjustment for the mediator was appropriate. Which effect is sought determines whether to adjust for the mediator or not.

As a rule of thumb, if the exposure is also a putative intervention target, it is the total effect that must be estimated. In biomedical research, adjustment for mediators is uncommon since the exposure is often a drug or a modifiable risk factor and is thus the target of intervention.

5. PARADOXES IN STATISTICAL MODELLING

Learning objectives

- Understand how apparent 'paradoxes' arise due to poor comprehension of causality
- Recognise the specific challenges with compositional data

When is statistical adjustment misleading?

There are situations where the 'correct' statistical adjustment in a linear model is not obvious, or indeed even tractable, and we look at some instances. The first is a problem that plagued the literature with confusion for decades and is an illustration of Simpson's paradox, and the second is an illustration of the challenges with compositional data.

The birthweight paradox: smoking during pregnancy and infant mortality

The birthweight paradox is famous as a 'paradox', even though there is nothing paradoxical from a causal framework perspective. It provides an excellent illustration of problems that stem from a limited comprehension of causal theory and subsequent misinterpretation of incorrectly specified multivariable models. We examine the association between **smoking** during pregnancy (exposure) and **infant mortality** (outcome) whilst 'adjusting' for **birthweight** (an alleged 'confounder').

A 'paradox' emerges because findings from the (misspecified) multivariable model are contrary to expectation, showing that:

- mean birthweight is lower amongst mothers who smoke during pregnancy compared to mothers who do not;
- overall infant mortality is *higher* amongst mothers who smoke during pregnancy compared to mothers who do not; whilst 'paradoxically',
- examining birthweight subgroups, infant mortality rates are *lower* amongst mothers who smoke during pregnancy than those who do not.

This was first exposed as a consequence of poor comprehension of causal inference by Hernandez-Diaz et al.[40] and Wilcox[41]. In October 2014, an entire edition of the IJE was dedicated to this topic. If data corresponding to this problem are categorised (Table 1), the phenomenon is recognised as Simpson's paradox, and if data are continuous and considered within a multivariable model (Table 2), the phenomenon is recognised more generally as the reversal paradox.

We illustrate Simpson's paradox with simulated data: one million mother and child pairs with data on birthweight, mothers' smoking behaviour during pregnancy, and infant mortality. Table 1 shows that rate ratios *within* birthweight groups is always <1.0 whilst *overall* (across groups) it is >1.0.

The reversal paradox is demonstrated in the multivariable regression model presented in Table 2, where the linear model that is not adjusted for birthweight yields *elevated* odds of infant mortality amongst mothers who smoke during pregnancy (OR = 1.07, 95%CI = 0.98-1.17), whilst the model adjusted for birthweight yields *reduced* odds (OR = 0.70, 95%CI = 0.64-0.77).

The problem lies in treating birthweight as a 'confounder'. Two potential causal relationships are given in the DAGs of Figure 7. There is evidence that lower birthweight children are more at risk of infant mortality due to causal antecedents that affect both foetal health (perhaps leading to premature birth) and infant health, thereby causing a greater risk of infant mortality. There is also

strong evidence that smoking during pregnancy causes lower birthweight, and is hence a descendant of the smoking exposure.

Table 1: Simulated data to illustrate the birthweight paradox: birthweight, mother's smoking behaviour during pregnancy, and infant mortality for 1 million mother and child pairs

Birth weight Range (Kg)	Mothers who smoked			Mothers who did not smoke			Rate Ratio
	Live Births	Infant Deaths	Mortality Rate[1]	Live Births	Infant Deaths	Mortality Rate[1]	
(0.5,1]	2	1	500.0				
(1,1.5]	64	2	31.3	68	6	88.2	0.35
(1.5,2]	1,394	30	21.5	2,250	59	26.2	0.82
(2,2.5]	10,360	127	12.3	30,018	524	17.5	0.70
(2.5,3]	30,318	188	6.2	158,876	1,453	9.1	0.68
(3,3.5]	36,694	143	3.9	329,896	1,528	4.6	0.84
(3.5,4]	17,406	26	1.5	275,228	692	2.5	0.59
(4,4.5]	3,510	3	0.9	91,288	102	1.1	0.76
(4.5,5]	244	0	0.0	11,768	12	1.0	
(5,5.5]	8	0	0.0	600	0	0.0	
(5.5,6]				8	0	0.0	
Total	100,000	520	5.2	900,000	4,376	4.9	1.07

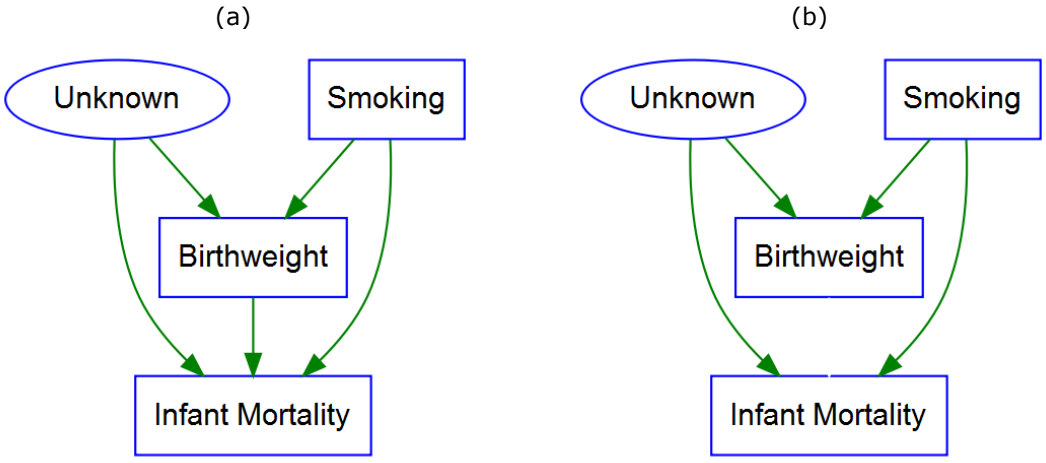
[1] per 1000 live births

Table 2: Regression model of infant mortality (outcome) on mother's smoking behaviour during pregnancy (exposure) both unadjusted and adjusted for infant birthweight for the simulated data summarised in Table 1

Model	Estimate	95% CI
Smoking exposure during pregnancy (unadjusted for birthweight)		
Non-exposed mortality rate[1]	4.86	4.72, 5.01
Smoking exposure odds ratio	1.07	0.98, 1.17
Smoking exposure adjusted for birthweight		
Base mortality rate[1,2]	3.32	3.19, 3.45
Smoking exposure odds ratio	0.70	0.64, 0.77
Birthweight odds ratio[3]	0.25	0.23, 0.26

[1] per 1000 live births; [2] centred on birthweight of 3.5 Kg; [3] per 1 Kg increase in birthweight

Figure 7: DAGs for the relationships amongst mothers smoking behaviour during pregnancy, their infant birthweight and risk of infant mortality, with common unknown causes of infant mortality and birthweight: birthweight (a) is or (b) is not causally related to mortality



Birthweight is a mediator in Figure 7a and therefore should not be adjusted for, as the effect sought is the *total* causal impact of smoking during pregnancy on infant mortality. However, if birthweight does not cause infant mortality (Figure 7b), since adjusting for it as a proxy confounder introduces a conditional relationship between smoking and unknown confounders; the antecedent unknown competing exposures (Figure 7b) become correlated with the smoking exposure, which will lead to

a biased estimate of the causal effects of smoking on infant mortality. In the language of graphical model theory, conditioning on birthweight invokes the 'back-door criterion' by unblocking the 'collider' that birthweight represents, partitioning the causal effect of smoking during pregnancy on infant mortality into direct (smoking-related) and indirect (due to unknown antecedent competing exposures).

Compositional data

It is important to remember that when we regress Y on X, adjusting for Z, we are asking: *what is the relationship of X with Y whilst keeping Z constant?* The assumption made is that the X-Y relationship is the same for all values of Z, i.e. the relationship is conditionally 'independent' of Z. We therefore need to think carefully about the implications of holding Z constant. For instance, what if $Z=X^2$, i.e. we have the quadratic model: $Y=\beta_0+\beta_1X+\beta_2X^2$? Clearly, we cannot interpret the coefficient for X (β_1) as though X^2 were constant; this instead requires the *joint* interpretation of the coefficients for X and X^2, i.e. β_1 and β_2 must be considered simultaneously when seeking to understand the X-Y relationship. This is perhaps not too challenging an issue if we are familiar with interpreting *curvilinear* relationships, but there are more complex scenarios that often go unnoticed where the same issue arises.

A particularly challenging scenario is when data are 'compositional', i.e. where constituent parts make up the whole. For instance, leg length is associated with human health[42], yet trunk length (including head) combined with leg length makes up total body height. Birthweight and weight gain combine to make current weight[26] and body size measures throughout life, from conception (zero weight) through birthweight to current weight are compositional data, since each *change* in weight adds to create current weight[43]. In nutritional epidemiology, statistical 'adjustment' within a regression model for total energy intake is often considered normal (and by some essential) when exploring health status in relation to constituent components of diet[44], yet there is limited appreciation that the constituents provide components of energy that add to make total energy intake. Research into the effects of physical activity seek to record periods of each day that individuals spend in sedentary behaviour, doing vigorous exercise, etc., including sleep, which all combine to make the complete 24-hour day[45]. Variations in one component *must* therefore impact upon variations in other components, since the length of the day is fixed; yet this is rarely, if ever, acknowledged.

Consider the example of height and its components leg length and trunk length (head included) in relation to risk of coronary heart disease[46]. Analysis of these variables in relation to any health outcome (e.g. blood pressure, BP) would involve regressing the outcome on each component (plus additional confounders). If an analyst wanted to investigate the association with leg length, for instance, they might wish to be sure that any association found was due solely to leg length and not because people with longer legs tend to be taller. Consequently, the analyst might decide to 'adjust' for height, *as though this were a confounder* (i.e. total height might be viewed as 'causing' leg length). There are concerns with this, however, since leg length is part of overall height, and we need to think carefully about what it means to increase leg length (the exposure) by say 10cm, whilst keeping height constant. This can only be achieved by decreasing trunk length by 10cm. Thus, the coefficient for leg length in such a model will be the difference in outcome between two hypothetical people with the same height, but where one has longer legs *and* a shorter trunk than

the other. An association could arise either because of an outcome-leg-length relationship or because of an outcome-trunk-length relationship.

If the analysis were repeated using trunk length as the exposure (retaining total height as the alleged 'confounder') the coefficient would be the same magnitude but opposite in sign. An alternative might be to regress the outcome on leg length and trunk length (which replaces total height), whilst retaining any other relevant confounders. The coefficient for leg length will then be the difference in outcome between two hypothetical people, where both have the same trunk length but one has longer legs (and is therefore also taller) than the other. An association could thus arise either because of a relationship between the outcome and leg length or because of a relationship between the outcome and total height. How do we distinguish between leg length and total height, as to which is more strongly associated with the outcome? Perhaps both are associated with the outcome? If true association is between the outcome and total height only, the coefficients for leg and trunk length would be equal.

This example shows how much care is needed in thinking about what it means to 'adjust' for a variable and illustrates the difficulty that arises in separating the effects of variables that are related structurally (e.g. height = leg + trunk). Structural relationships amongst variables in regression models are ubiquitous yet often overlooked; we cover this again when we discuss mathematical coupling (MC) and its impacts in *analysis of change* and use of *ratio variables*. Understanding causal relationships amongst variables explored in a multivariable regression model is central to the interpretation of that model. Model development should therefore be driven by *a priori* understanding of causal relationships. One can use DAGs to set out a view of causal relationships amongst relevant variables (recorded and not recorded). Researchers are responsible for making appropriate decisions regarding their assumptions reflected in their DAG. Assumptions must be explicit else unintended *implicit* assumptions may arise. As already stated, for instance, the absence of arcs in DAGs represents important assumptions that can sometimes be quite strong. Choices made in a DAG may substantially alter the variables nominated for inclusion in regression models, which in turn has the potential to alter model findings and subsequent interpretation.

For compositional data, it can often be challenging to affirm with any degree of certainty whether components cause the whole or the whole causes the components (e.g. do leg length and trunk length combine to *cause* total height, or does total height *cause* both leg length and trunk length?). This dilemma and its associated implications are examined again when we look at compositional data in the context of ratio variables: we show that understanding context to inform the 'correct' model is crucial though far from straightforward, and sometimes it may be impossible to determine a 'correct' model!

Summary

There are many instances in observational research where multivariable models are employed and focus is on estimated exposure effect size interpreted to be the **total causal effect** of the exposure on the outcome. If the adjustment set of covariates employed in the multivariable model is not carefully and robustly justified within a causal framework, estimated effect sizes may be seriously misleading; there may be no optimal subset of observed covariates that allows for robust causal inference in some instances. Failure to work within a causal framework gives rise to considerable misunderstandings in the literature.

6. CONDITIONING ON THE OUTCOME

Learning objectives

- Understand the broad implications of regression to the mean (RTM)
- Learn how conditioning on the outcome introduces statistical artefact due to RTM
- Be aware of how investigation of longitudinal data can become conditional on the outcome
- Recognise how to view longitudinal data correctly within a causal framework

Regression to the mean (RTM)

Most statisticians will likely have come across regression to the mean (RTM) and many therefore believe they understand it. However, the effects of RTM are more widespread than typically appreciated[47-51]. This is in part due to the narrow way in which the concept is usually taught, though also because RTM can operate in ways that are not obvious and are easily overlooked. Although RTM is typically attributed to measurement error, it is not necessarily 'error' *per se* that is key (which would often dilute the effect size and widen standard errors); all other occurrences of RTM are poorly understood. It is therefore important to examine RTM thoroughly to appreciate its consequences, especially in the context of multivariable statistical modelling.

In providing a definition of RTM, most textbooks refer to the phenomenon where a variable, if extreme on its first measurement, tends to be closer to the centre of the variable distribution on a subsequent measurement. This oversimplification can lead to the incorrect view that RTM occurs only across repeated measures of the *same* variable. The implicit variation described is also most often attributed to measurement 'error', which is misleading.

A better definition is that, **following an extreme *random* event, the next *random* event is likely to be less extreme**. The concept of 'random' does not involve 'error' of any kind. For instance, take two independent normally distributed variables, X and Y (i.e. their correlation is zero). If, in one instance, we have a value of X far from the mean of X (i.e. an unusually high value for X), we are more likely to have a value for Y that is closer to the mean of the distribution of Y, since $E(Y|X) = E(Y)$. RTM may thus arise for measures that are not repeats of the *same subject*, nor even the *same variable*, and this does not need to involve any form of *error*. Hence, the scope for RTM is enormous, and it is this that is overlooked.

Sir Francis Galton (right) illustrated that RTM can occur across measures that are not repeats of the same subjects when he described RTM for the very first time in 1886[52]. Galton collected self-reported body heights for families. As men are on average taller than women, women's heights were multiplied by 1.08. For each set of parents, Galton plotted the average of the parents' heights (he called it mid-parent height) against the heights of their offspring.

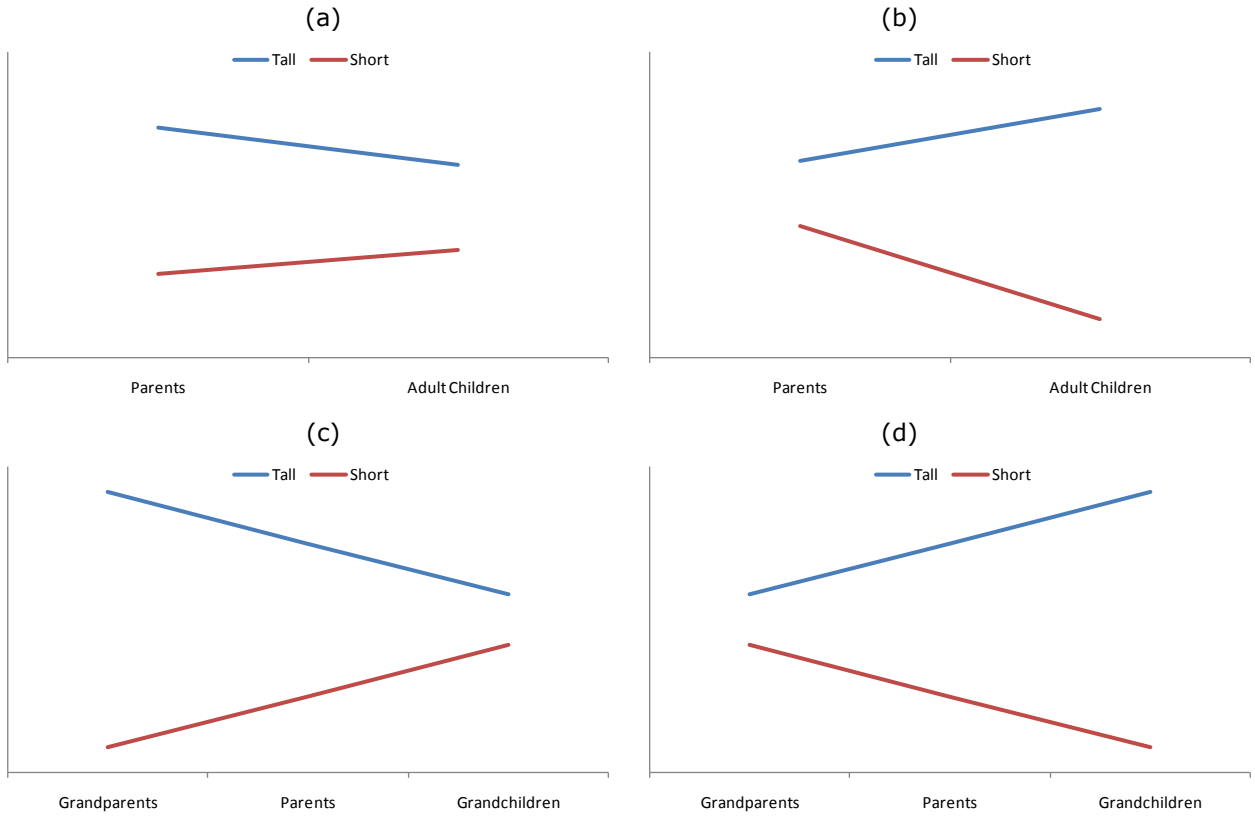
Although adult children of tall parents were taller than most, they were, on average, shorter than their parents. In contrast, adult children of short parents, whilst shorter than most were, on average, taller than their parents.

Clearly, **the same subjects were not involved in the repeated measures**, as successive generations were being assessed. Figure 1a shows the trend of body heights across the two generations, grouping families according to parents' heights by defining parents as 'tall' (≥68 inches) or 'short' (<68 inches). The mean heights of 'tall' and 'short' parents were 69.51 and 66.66

inches, respectively, whereas the mean height of adult children from tall parents was 68.79 inches and of adult children from short parents was 67.12 inches. Human heights appeared to converge across generations, **suggesting that after many generations there would be fewer very tall or very short people**.

This is where most standard texts stall at recounting this story, leaving readers only seeing half of what is effectively a two-part story. Using the same data, by grouping the families according to adult children's heights (as opposed to parents' heights), Figure 1b shows an apparently contradictory trend of heights across the two generations. Children were defined as 'tall' (≥68 inches) or 'short' (<68 inches), with mean heights in these groups of 69.89 and 65.77 inches, respectively. The mean height of parents of tall children was 68.87 inches and of short children was 67.58 inches. Heights then appear to diverge across the two generations, **suggesting that after many generations there would be more very tall or very short people**.

Figure 1: Trends in body height across generations: (a) mean body height of subgroups in Galton's data when the families are grouped by parents' height; or (b) by children's height; (c) trend in body height across three generations when families are grouped by grandparents' height; or (d) by grandchildren's height



Neither contradictory interpretation of Galton's data is correct. Patterns in Figures 1a & 1b are consequences of the correlation between heights of parents and heights of children being imperfect (i.e. <1). RTM occurs when children's heights are regressed on their parents' heights, or vice versa, in the presence of a *less-than-perfect correlation* between both height variables.

The same phenomenon can thus arise for regression between any two variables with less-than-perfect correlation; hence, **all regression suffers RTM** as there rarely exists perfect correlation between any two variables. Furthermore, **RTM is not limited to regression**. Campbell and Kenny pointed out that any factor that makes the correlation of two variables less than perfect can cause RTM[53]. We return to these points later.

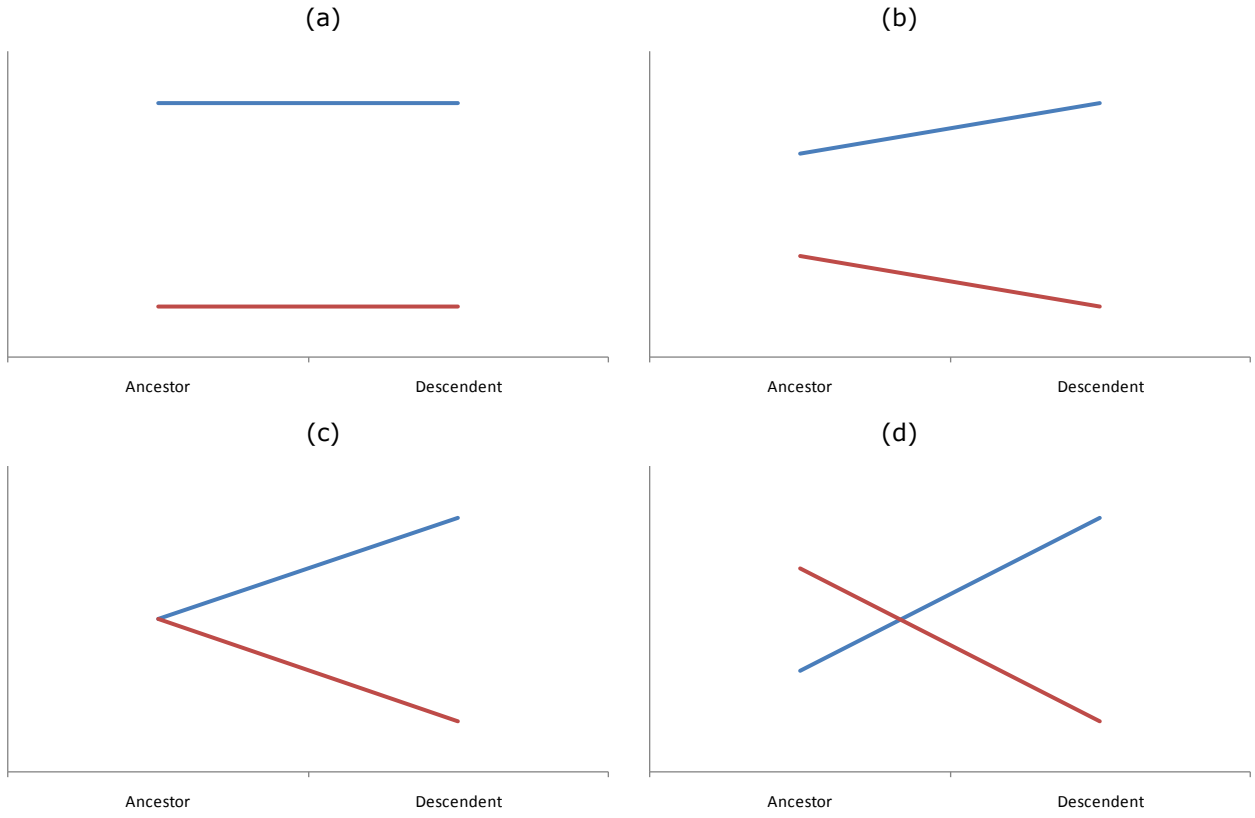
RTM and longitudinal data

If Galton had obtained records of body heights of grandparents for his families, and if families were grouped according to grandparents' heights, trends would look like Figure 1c; if families were grouped according to the grandchildren's heights, trends would look like Figure 1d. Convergence or divergence becomes more notable than in Figures 1a & 1b since the correlation between the heights of grandparents and grandchildren is smaller than between grandparents and parents, or between parents and children.

If there was a long historical record of heights for these families, and families were grouped according to heights of the latest generation, the trend of body heights back in time would converge. The positive correlation between heights of successive generations becomes smaller the farther back the genealogy goes.

Figure 2 generalises this for the relationships between ancestors' and descendants' heights, where data are grouped according to whether the descendant's height is above or below the sample mean. If the correlation between ancestor and descendant height is perfect and positive (i.e. $\equiv 1$), the two lines are parallel (Figure 2a). When the correlation lies between 0 and 1, heights appear to diverge looking forwards in time (Figure 2b). If the correlation is 0, there is no difference in mean ancestral height between the two groups (Figure 2c). If the correlation is negative, the two lines must cross (Figure 2d).

Figure 2: Representation of the generalisation of RTM between ancestor and descendant body heights: (a) correlation = 1; (b) $0 <$ correlation < 1; (c) correlation $= 0$; (d) correlation < 0.



RTM in regression analysis

One might anticipate that RTM occurs within regression, given the word 'regression' is used for both. Indeed, when one variable is regressed on another, and both are imperfectly correlated, RTM is inevitable. If this is due entirely to measurement error, this is known as ***regression dilution***.

However, RTM should not be attributed to measurement error alone, with no regard for other potential factors (e.g. confounding, variation in physiological response, or exogenous factors that lead to imperfect correlations amongst the variables of interest). Consider, for instance, systolic blood pressure (SBP), which is a clinical outcome that is renowned for its lack of precision, due in part to measurement error though more often due to confounding or physiological variation in response (e.g. 'white coat hypertension'[2]). This clinical measure provides a good illustration of the inherent underlying biological or physiological variation that can play an important yet often overlooked role in the occurrence of RTM.

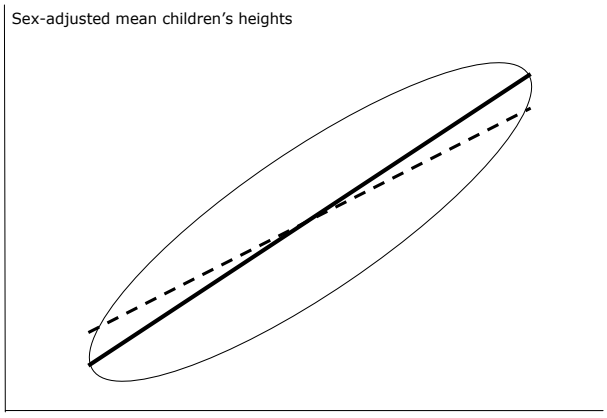
The role of biological / physiological variation

Imperfect correlations often arise due to biological or physiological variation, or both[47-51]. For instance, one is likely to obtain different readings of systolic blood pressure (SBP) for the same individual, even when a series of measures are made over a relatively short time period. This may be attributed to the device used (or the person who uses the device) not being entirely reliable (i.e. measurement error); more likely, however, the underlying 'true' SBP (i.e. assuming blood pressure could be measured error-free) naturally fluctuates around a mean value. Fluctuation may be inherent due to biological (genetic) or physiological (environmental) factors, and might be short-term (seconds), medium term (years), or long-term (across generations).

The latter was observed in Galton's data, with the heights of children regressed on the heights of their parents (Figure 3). Data points form an ellipse around the axis of equality (solid line: $y = x$) due to imperfect correlation between parents' and children's heights, which is largely down to biological variation in heights across generations, not measurement error.

Placing parent heights on the *x*-axis and offspring heights on the y-axis (Figure 3), mid-parent height regressed on the heights of their adult children yields a slope <1. If the heights of adult children were regressed on mid-parent heights, the slope is >1. Thus, there are two forms of regression: for the 1st the model is conditional on parents' heights (dividing data as in Figure 1a); for the 2nd the model is conditional on children's heights (dividing data as in Figure 1b).

Figure 3: RTM where the mean height of children is regressed on the mean height of parents



There are two forms of regression due to RTM. It is less important to distinguish the two forms of regression than to recognise the role of RTM on the estimation of the regression model coefficients (i.e. slope, in Figure 3). In multivariable regression, the joint estimation of multiple covariates, each with an imperfect correlation with the outcome, can lead to many strange consequences due to RTM that are poorly understood.

Conditioning on the outcome

If we recruit patients because they exhibit disease (e.g. hypertension), there is a danger that we select merely an extreme realisation of a naturally (randomly) varying phenomenon (hence the selection of a random variable concept is introduced again). Whilst the body seeks to maintain

appropriate levels of systolic blood pressure (SBP), actual SBP levels vary a lot quite naturally (exogenous factors may be operating which might explain some of this variation, but for simplicity we view the variation as effectively 'random' noise).

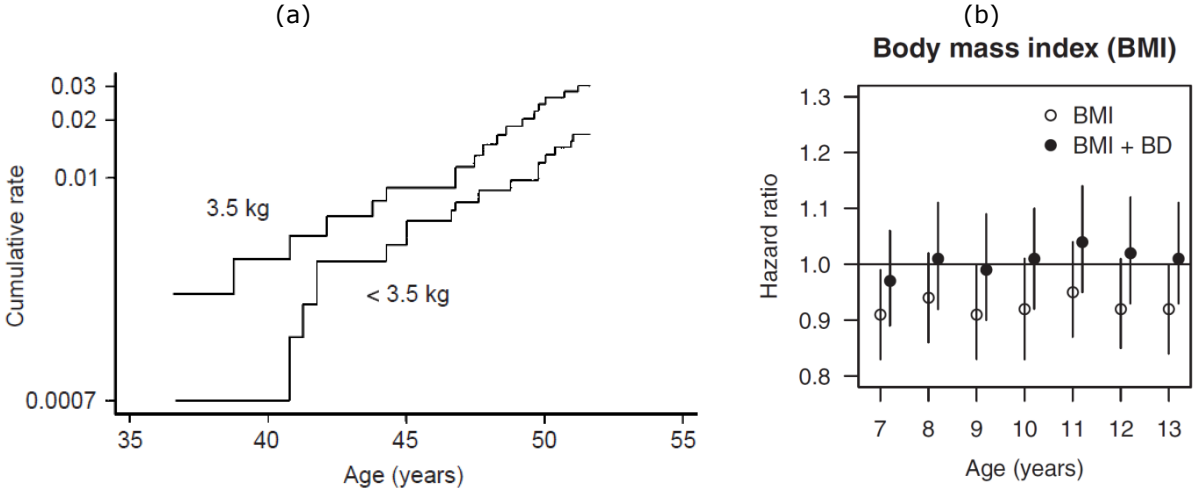
If we took 1000 individuals and recorded their SBP to select those with SBP values greater than a threshold to yield, say, ~10% of our sample, and if this subsample were evaluated the next day, on average they would demonstrate a lower mean SBP. Nothing unusual has happened; this is merely a consequence of RTM where conditioning is on the (randomly varying) outcome. Hence we need control groups and random allocation to treatment within RCTs: to contrast change that will occur anyhow to that which is amplified under the influence of an intervention.

Generally, if any *conditioning* occurs in the evaluation of imperfectly correlated variables, RTM will emerge. Moreover, its effects may be hard to spot, and its magnitude unknown. We illustrate this for lifecourse research, evaluating longitudinal body sizes in relation to a later-life health outcome.

Growth 'trajectories'

Implicit conditioning on an outcome may lead to the (mis-)interpretation of what erroneously is termed a lifecourse 'trajectory'. For instance, in the evaluation of body size throughout the lifecourse in relation to the risk of developing breast cancer in later life[54-56], implicit conditioning is on later-life disease status. One might see a 'pattern' of risk in relation to birthweight[55] (Figure 4a) and in relation to body mass throughout early life[56] (Figure 4b), from which it might then be argued that higher birthweight is associated with *increased risk* of breast cancer[55], while greater body mass index (BMI) at ages 7-13 is associated with *reduced risk* of breast cancer. Intriguingly, the latter is mitigated (i.e. the protective effect diminishes) if 'statistical adjustment' is made for breast density (as assessed by a mammogram)[56].

Figure 4: (a) Cumulative breast cancer incidence rates by age and birthweight (reproduced from[55]); (b) Association hazard ratios (HRs) and 95% confidence intervals by age between breast cancer and body mass index unadjusted and adjusted for breast density (reproduced from[56]).

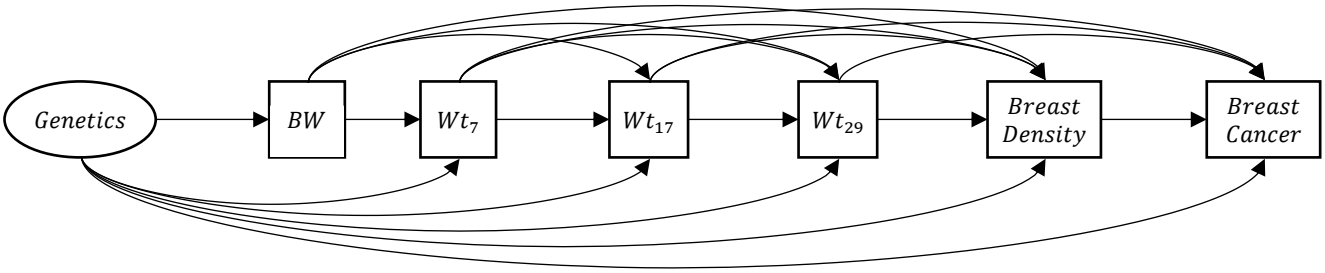


We examine this scenario in a causal framework and draw a DAG (Figure 5); the total causal effect of birthweight on cancer risk cannot be estimated by covariate adjustment, as we do not have any measured genetic data. Furthermore, body sizes after birth and adult breast density are mediators of the birthweight-outcome relationship (introducing the risk of reversal paradox if adjustments are made for intermediate body sizes). Some might argue that breast density is an appropriate 'proxy confounder' for the unmeasured (maybe unknown?) genetic effects, i.e. a similar argument

to that of adjusting for current body size in the birthweight-blood-pressure relationship. However, the DAG indicates that **statistical adjustment should never involve breast density**, which points to the approach adopted in Figure 4b as inappropriate.

Even if the genetics information were known, attributing any meaningful interpretation to the impact of each BMI variable is challenging. For BMI age 7, for instance, adjustment involves only birthweight; for older BMI exposures, adjustment involves all preceding BMI measures. Yet again, we have multiple models, each indicating the total causal effect of body size at specific ages, with

Figure 5: DAG of unobserved (latent) genetics and observed birthweight and BMI several ages up to age 29 in adulthood, breast density, and breast cancer.



no overall 'holistic' take on the impact of a 'growth trajectory', no summary synopsis of how growth throughout the lifecourse affects the risk of breast cancer.

The big challenge

It is a holistic approach that some have sought to address, though introducing more problems than have been solved to date. How to both disentangle and yet overall summarise the causal effects of a time-varying exposure (e.g. body sizes) on a later-life outcome remains challenging. Methods for lifecourse research data are constantly under development, some not yet published while others only recently so[57]. With a paucity of well-developed methods, spurious approaches have been made to investigate lifecourse data; some have reverted to simple graphical display of their data, from which inference is drawn relying heavily only on intuition, which unfortunately overlooks RTM. Two examples appear in the *New England Journal of Medicine*[58;59]. To understand these, we must first understand z-scores, as used by these publications.

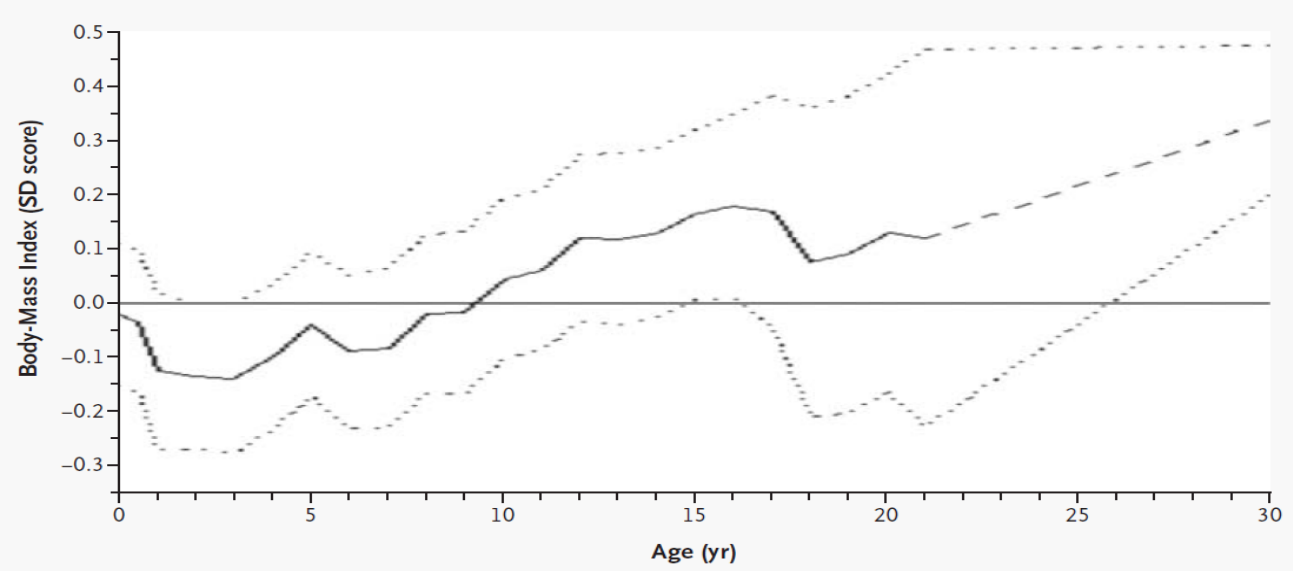
Z-scores

The standard score is the signed number of standard deviations by which a value of an observation or data point is above or below the mean value of an observed measure: values above the mean are positive, while values below the mean are negative. The standard score is a dimensionless quantity obtained by subtracting the population mean from an individual raw score and dividing by the population standard deviation. This process is called standardising or normalizing (not to be confused with the use of 'normalizing' that refers to types of ratios; see later lectures) and standard scores are also called SD-scores, z-scores, z-values, and standardised variables. They are most frequently used to contrast an individual's measure to the population standard distribution (which need not be normal, though often normality is assumed and they are then also known as 'normal scores'). Computing a z-score requires knowing the mean and standard deviation of the population from which a data is sampled; where only the sample is available for estimates of the population mean and standard deviation, the standard score yields the Student's t-statistic.

The false 'trajectory': misinterpretation of a graphical display

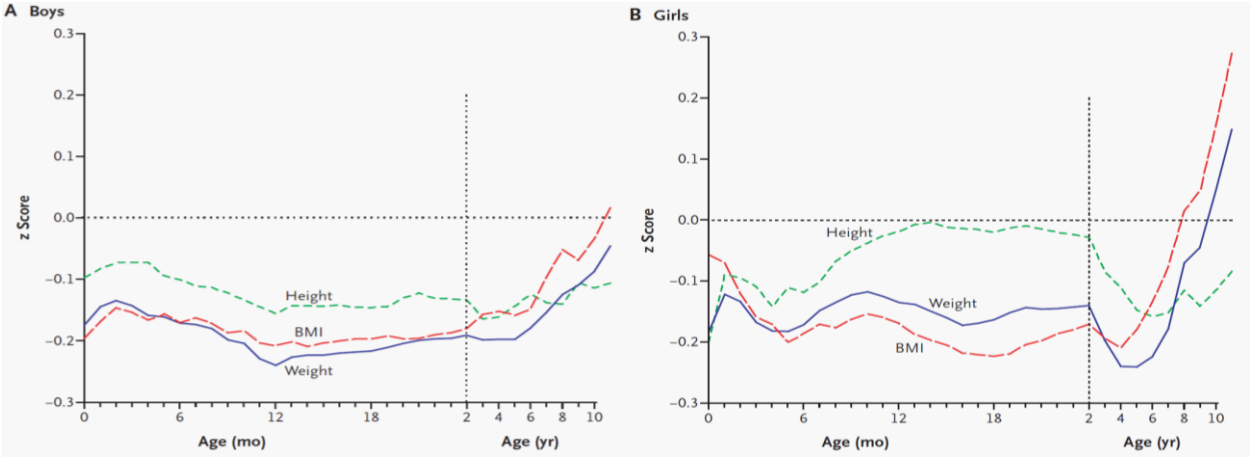
One NEJM article examined impaired glucose tolerance (IGT) amongst men and women aged 26-32 years in relation to their standardised (i.e. SD-score) BMI throughout life, plotting the results for those **with IGT/diabetes** (Figure 6)[58]. From graphical presentation only, the authors state that individuals with IGT/diabetes "*had a low body-mass index up to the age of two years, followed by an early adiposity rebound (the age after infancy when body mass starts to rise) and an accelerated increase in body-mass index until adulthood*", from which they conclude: "*crossing into higher categories of body-mass index*" after age two is "*associated with these disorders*"[58].

Figure 6: Body mass index by age for individuals with impaired glucose tolerance or diabetes developed; reproduced from[58]



Another NEJM article, entitled "*Trajectories of growth among children who have coronary events as adults*"[59] charts childhood growth for 8760 people born in Helsinki in 1934-1944. It was noted that **amongst those who had had an adult coronary event**, they had been, on average, small at birth, thin at age two, and thereafter rapidly put on weight (Figure 7). The authors conclude: "*the risk of coronary events is more strongly related to the tempo of childhood gain in body mass index (BMI) than to the BMI attained at any particular age*."

Figure 7: Mean z-scores for Height, Weight, and Body Mass Index (BMI) in the first 11 years after birth among (A) boys and (B) girls who had coronary heart disease as adults; reproduced from[59]

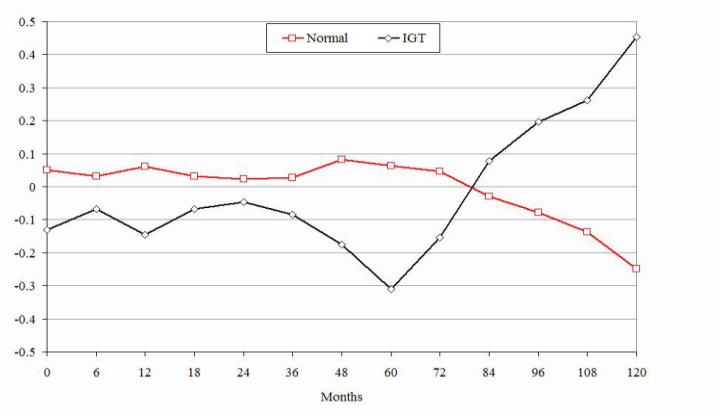


In both articles, selection was made of ***individuals already suffering the condition of interest***, for which investigation was then made of prior "growth trajectories". What is missing from Figures 6 and 7, that helps understand the problem, is corresponding ***data on non-affected individuals***. Similar data were made available for a limited period on the DOHaD conference website.

Data show patterns anticipated for affected *and* non-affected individuals, examined using BMI standardised scores at each age (Figure 8).

If the groups with and without disease were identical in size, the resulting pattern for non-affected individuals would be a mirror (about the zero-axis) of the pattern for affected individuals, since standardised scores sum to zero at each time point. With more non-affected individuals, this group mirrors the pattern for affected individuals with smaller amplitude.

Figure 8: BMI SD-score by age for individuals with (black) or without (red) impaired glucose tolerance (IGT)



Note: There appears a data coding error around 60 months, as the mirroring is not correct.

It is widely accepted that cardiovascular disease, IGT, and type 2 diabetes are linked to obesity[60], and the strongest association with body size (e.g. BMI) occurring prior to and concurrent with the onset of disease. Therefore, this information, along with the similarity of Figure 8 (and Figures 6 and 7) to Figure 2d, outlines the critical issues:

- **Graphical summary** of such data is **conditional on the outcome**, since data are divided according to the outcome.
- Dividing data on the outcome is tantamount to dividing data on the most recent exposure measure, due to their strong association / correlation.
- Inferences from articles **derived only from inspection of the graphical summary of data** divided this way[58;59] yield the same invalid, contradictory interpretations as Galton's data.

So-called body-size 'trajectories' by age for diseased and non-diseased groups 'appear' to converge when looking back over age, just as in Figure 2d where lines joining data points crossed. Although we cannot dismiss potential associations between later-life outcomes and "*early adiposity rebound*" or *rapid compensatory growth at earlier ages*, we must recognise that such patterns, as in Figures 6-8, arise due to RTM and conditioning on the outcome *even with no genuine causal link*.

To interpret Figures 6-8 correctly, recall that each data point is a bivariate correlation between the later-life outcome and exposure at each assessment occasion. Figures 6-8 thus display a series of bivariate correlations. It is not at all apparent why these data points should be joined by a line, as it gives the erroneous impression of an underlying dynamic, from which one might mistakenly infer that the *relative* position of each data point has meaning in terms of *change* in exposure associated with the outcome. The joining of bivariate corrections with lines and using the term 'trajectory' is seriously misleading. Figures 6-8 only state that:

- Birth size is ***negatively*** associated with the IGT and diabetes – a familiar finding attributed already to birthweight in relation to cardiovascular diseases and related conditions (the association is reversed for cancer[55]).

- Adult (current) body size is ***positively*** associated with the IGT and diabetes – a familiar finding attributed already to adult obesity in relation to cardiovascular diseases and related conditions (the association is reversed for cancer[56]).
- Intermediate body sizes must be associated with cardiovascular diseases (and cancer) with a pattern that interpolates between the relationships for birth size and adult body size (not stated explicitly previously, but this is implicit already).
- The correlation between body size and outcome (cardiovascular or cancer) must be ***zero*** at some intermediate age (a statistical consequence with uncertain clinical implications).

Focus is sometimes given to the age at which the 'crossing' of z-scores occurs, as though this were a 'critical' development period with **clinical meaning**. There may be no causal link between the later-life outcome and body size at the age of crossing z-scores, since the data summaries in Figures 6-8 are consistent with influences on the later-life outcomes arising only from birth size (for all that it bestows via in-utero 'programming'[61;62]) and adult body size immediately prior to the outcome occurring; i.e. there need be no attributable influence from body sizes in between birth and adult body size, beyond the latter being established by growth throughout the lifecourse. The influence of higher-than-average body sizes throughout life might cause nothing more than over-weight just prior to the onset of cancer.

Summary

Assertions of causality inspired by graphical summary of correlations conditioned on the later-life outcome are erroneous. Statistical analyses that formalise such data summaries suffer RTM, with the magnitude of adverse impacts unknown. Despite warnings against such practices[63], they continue unabated[64].

Generally, conditioning on an outcome is potentially dangerous from the perspective of statistical inference. More generally, conditioning in any statistical analysis must be considered carefully. Within multivariable regression, conditioning is implicit and should be framed within a causal framework where causal inference is sought.

7. CONDITIONAL DATA ACQUISITION

Learning objectives

- Understand the causal implications of implicit conditioning in data acquisition or selection
- Appreciate the potential pitfalls of 'routine' data without due consideration of its provenance

Implicit conditioning

We have seen that exogenous (implicit) conditioning on an outcome (i.e. by sub-setting data based on outcome values) impacts severely on the robustness of causal interpretation sought from a multivariable model when association is sought between longitudinal exposure data and a later-life outcome. Conditioning on the outcome is explicit, as only individuals with the condition of interest are selected for analysis. However, sample selection more generally can be influenced implicitly, in ways not immediately apparent.

As statistical inference is generally focused on sample selection and associated sample variability (which is why we use measures of confidence to frame estimated relationships subject to sampling variation), study design and data acquisition more generally are critical to ensure robust statistical and causal inference. In addition to requiring that a sample is 'representative' of the population it purportedly represents, it is similarly vital that there are no statistical artefacts generated because of the data acquisition process – this last point is often overlooked and depends heavily on study design and conduct. This issue will only become a growing problem in the big data era, as data acquisition and its provenance are not as carefully considered as with traditional epidemiological study designs, for instance; most observational studies are unlikely to be as stringent in processing data as say the case-control or cohort study designs tend to be.

Looking forward, we are likely to be engaged in the analysis of increasing volumes of observational data that are collated through all kinds of acquisition processes; sample selection processes will be increasingly complex and, in some instances, not fully understood. It is in these instances that we need to appreciate the role of sample selection and the implicit conditioning that this context brings. We examine a situation in which the cavalier consideration of data selection is a potential root cause of complete lack of validity in analyses that follow.

Geographical analysis of disease incidence

To illustrate the importance of carefully reflecting upon the sample selection process and potential implicit conditioning, we examine research that investigates the link between population mixing and childhood leukaemia, as this has generated equivocal and contradictory results, perhaps due to inadvertent conditioning on the outcome[65]. Any analytical approach to this research question uses predefined geographical clusters to quantify the exposure (i.e. population mixing) and a measure of the outcome (i.e. number of childhood leukaemia cases); common to both exposure and outcome is the area population in which each measure is recorded.

The population mixing hypothesis proposes that the immune systems of children resident in more isolated and/or less densely populated communities are likely to have been exposed to a less diverse range of infectious agents than those resident in less isolated and/or more densely populated communities; and that such children are therefore more likely to develop leukaemia once exposed to novel infections transmitted by inward-migrants from elsewhere[66].

There are two principal analytical approaches considered to examine the relationship between population mixing and childhood leukaemia: (i) selecting areas according to specific characteristics and comparing the incidence of childhood leukaemia in these areas to that expected on the basis of the national average; or (ii) deriving a multivariable regression model of region-wide data to model characteristics associated with the incidence of childhood leukaemia.

Note: Were region-wide data used to create two ratio variables – one for a measure of population mixing per capita and one for the number of childhood leukaemia cases per capita – these two ratios would be mathematically coupled and their analysis by simple correlation would comprise both a spurious component (due to MC) and a genuine component (if any true association existed). Fortunately, in this instance, neither the sub-region or region-wide analytical strategies invoke MC using correlation or regression, as neither analyse the two ratio variables directly. The sub-region analytical strategy is selective of areas according to one or more *observed* ratio variables and contrasts this to the national *expected* ratio variable, but the analysis does not involve correlation or regression explicitly. The region-wide analytical strategy uses Poisson regression to calculate the *partial correlation* between the numerator (number of leukaemia cases) and exposures (population measures of population mixing), whilst 'adjusting' for population counts (typically logged and set as the 'offset' in Poisson), as proposed by Pearson, Neyman and Fisher[67-69].

Issues with the 'sub-region' analytical strategy

Some studies that used the 'sub-region' approach were instigated due to the apparent 'cluster' of leukaemia cases in an area[70], which aims to verify an 'excess' of cases, rather than testing against the null hypothesis, resulting in endogenous selection bias[71]. In some instances, it is unclear how such specific proxies for population mixing were chosen and thus difficult to determine whether these areas were also selected for investigation due to clusters, or suspected clusters, of cases. Examples of these specific population mixing proxies include the influx of servicemen to an area[72] and migration due to forestry developments[73]. This makes these studies difficult to reproduce and compare, and has resulted in the use of a wide variety of time frames being considered.

Often, several distinct time frames (or time frames combined) are investigated within a single study and it has been suggested that a deficit of cases in a time period immediately after a period of excess is a result of the suggested leukaemia-causing agent being mainly immunising after an epidemic[74]; by choosing time frames this way, they can be manipulated to show greater or smaller excess cases, but apparent deficits after a period of excess could also be a result of regression to the mean[50]. Where studies report an 'excess' of cases, this is not consistently in the same age group and often a single study will report on multiple age groups as well as age groups combined (where the 'excess' may or may not remain); these studies often include multiple tests, introducing associated issues.

Methods for evaluating the two approaches

The issues with the 'sub-region' approach may seem clear when stated, yet the population mixing hypothesis is founded on studies performed this way, whereas the 'region-wide' approach does not consistently support many 'sub-region' findings. It is therefore worth examining these two principal analytical strategies from a statistical standpoint in order to evaluate the robustness of evidence underpinning the population mixing hypothesis. We did this using simulations under the null hypothesis, i.e. whereby only the size of the population drives the incidence of cases. This means

that we evaluate the type I error rate (i.e. the rate at which false positive results are generated) and assess bias and flaws inherent to each approach. Simulations were informed by real-world data (i.e. using the correlation structure and approximate distributions of the variables in a real dataset) and the use of proxies for 'population mixing' that were introduced in the first study on the topic[66]: *population density* (all those capable of spreading a leukaemia causing agent); and *inward-migration* (the relative number of new arrivals capable of bringing such an agent with them, expressed as the proportion of migrants within the population):

- **Population Density** = Total Population/Area (km^2);
- **Proportion of Inward-Migrants** = Inward-Migrants/Total Population.

Simulations were also compared to the analyses of the real-world data in which, according to the hypothesis, an association should exist between population mixing and childhood leukaemia.

The 'sub-region' approach selects areas for analysis (in our example, electoral wards) based on extreme values of the population mixing proxies (*low* population density and *high* proportion of inward-migrants) or *high* incidence of childhood leukaemia compared to that expected given the national average. We generated 15 scenarios in which areas for analysis are selected according to all possible sequences of these three variables. The 'region-wide' approach generated Poisson regression models of childhood leukaemia incidence with population density and/or proportion of inward-migration as covariates.

Findings from a simulation

Based on 10,000 iterations of the simulation and corresponding analyses on the real-world data using the 'sub-region' approach we show that, in randomly selecting wards with high incidence rates solely or in combination with other characteristics, an overall higher than expected incidence of childhood leukaemia is consistently observed (Figures 3 and 4). In contrast, results drawn from the 'region-wide' approach conflict, suggesting a 'protective' effect of high inward-migration and a 'detrimental' effect of low population density of childhood leukaemia incidence (Figure 3).

We demonstrate the problem of 'targeted' selection, which arises from implicitly conditioning on the outcome, i.e. where attention is given to areas with a high incidence of childhood leukaemia and these areas are subsequently selected to evaluate the association of childhood leukaemia with population mixing. The problem arises because it is difficult to be sure that selection according to the choice of population mixing characteristics is not affected (subliminally) by some knowledge of the outcome. The problem is exacerbated in this case due to the apparent clustering of the outcome within geographical areas, as when there is variation in the size of areas; in such instances, there is inevitably larger variation in the occurrence of cases. It would therefore be expected that the variation in leukaemia incidence is greater in areas with small populations and smaller in areas with large populations and there need be no causal explanation for apparently high incidence rates in areas with small populations; this can merely be the result of sampling variation[75;76]. It is human fallacy to overlook this and instead naturally seek causal inference[77]. If there appears to be a high incidence of cases in a particular sub-region over a specified time period, this could draw attention to the kind of misleading analyses outlined here, whilst this high incidence may not be maintained in any subsequent time period. Analyses on region-wide data (whole or randomly selected), on the

other hand, guarantees that conditioning on the outcome is avoided, as too therefore are any consequent biases we have highlighted for the sub-region analyses.

Figure 3: Percentage of statistically significant results at the 5%-level by strategy for both simulated and observed data. 'Subsample' selection strategy results were analysed using the binomial exact test, and the direction of the bars indicate whether the estimated probabilities of the significant test results were greater (>0) or less than (<0) the national average. 'Region-wide' strategy results were analysed using Poisson regression, and the direction of the bars indicate whether statistically significant coefficients were greater (>0) or less than (<0) the null of zero.

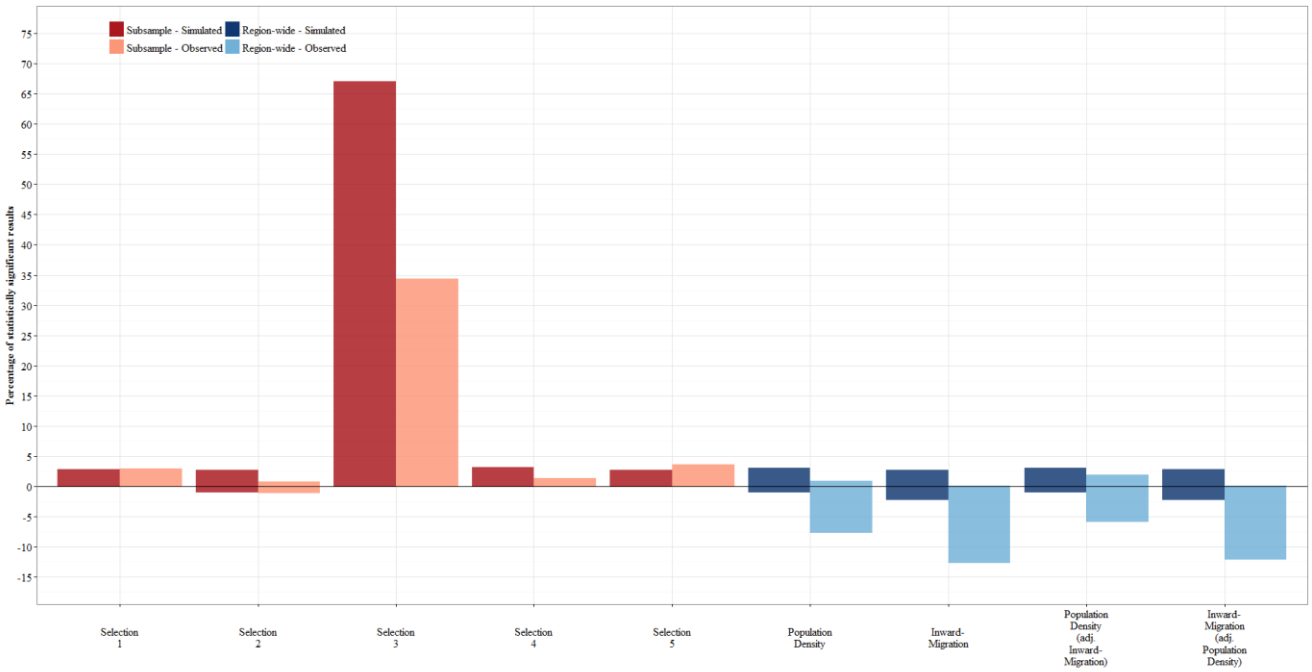
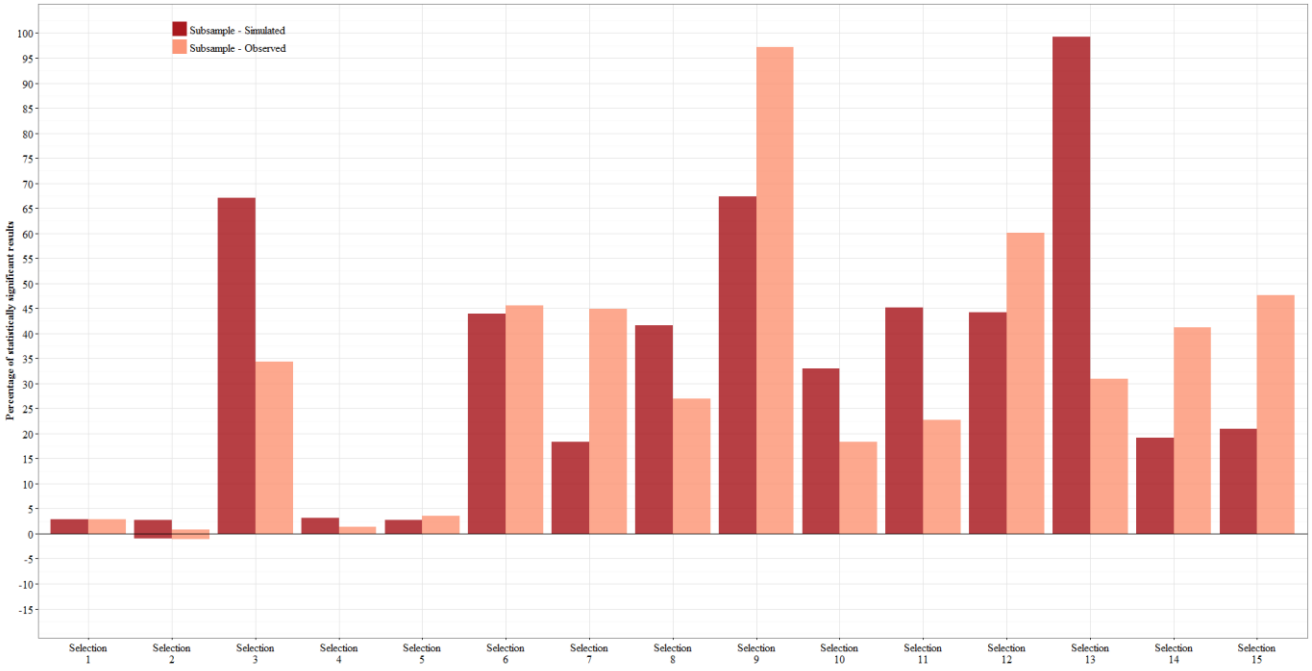


Figure 4: Percentage of statistically significant results at the 5%-level using the 'subsample' selection strategy for both simulated and observed data; direction of bars indicates whether the estimated probabilities of the significant test results were greater than (>0 on graph) or less than (<0 on graph) the national average. Pop Den = population density, In-Mig = inward-migration, Inc = childhood leukaemia incidence.



The substantive contrast between findings that condition on the outcome and all other analytical strategies illustrates that a genuine *negative* effect might be erroneously reversed due to the wrong

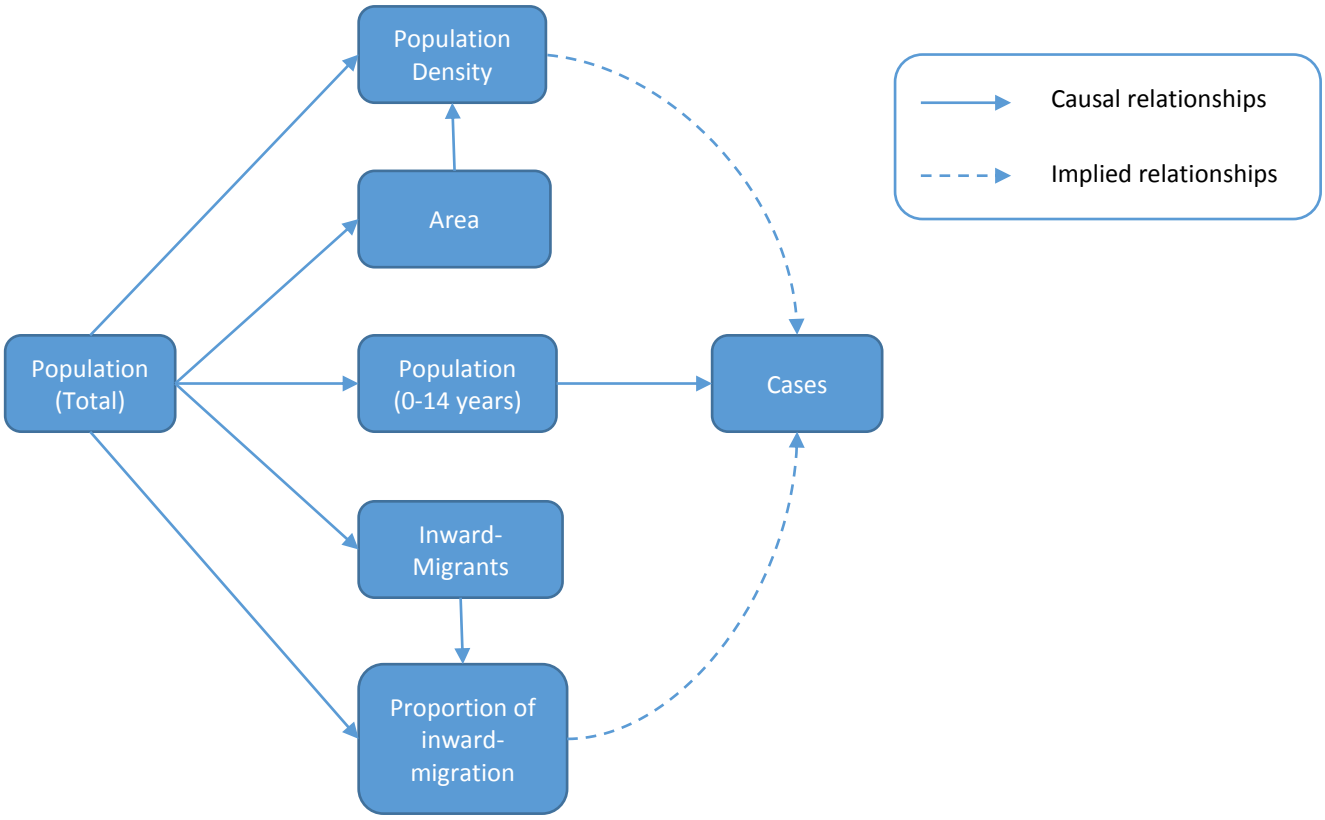
analytical approach. Combined with the risk of publication bias, where disproportionate attention is given to positive findings, much of the existing literature that conditions on the outcome when examining 'population mixing' would **severely skew the 'evidence' in favour of a positive effect if this were not true, or even if the opposite were true**.

It has been reported that there is evidence of publication bias[78] and there is no way of knowing when manuscripts have been rejected due to showing a null (or opposite) association. If the entire dataset for a region is unavailable, sampling of small areas must be random and region-wide to avoid (subliminally) conditioning on the outcome of interest. Failure to adopt a 'region-wide' analytical strategy for investigating the association between childhood leukaemia and population mixing will likely yield false findings.

Mathematical coupling (MC) amongst model covariates

Mathematical coupling (which is covered in detail in Chapter 9) arises for the geographical analysis of the population mixing hypothesis, which is why analyses on region-wide data (whole or randomly selected) reveal a modest bias: i.e. the small but notable skew towards higher than expected childhood leukaemia incidence (i.e. more significant *p*-values than expected). This is due to the regression model including mathematically coupled ratio variables as covariates: a small but non-zero bivariate correlation arises between the outcome (number of cases) and each exposure (population mixing ratio variable) due to the confounder (total population) being involved in the construction of each exposure variable (Figure 5).

Figure 5: Graph representing the simulated relationships of the dataset. Causal relationships are represented by solid arrows and implied causal relationships are represented by dashed arrows.



Summary

It is human nature to go with what 'feels right' and not to question what at face value supports or coincides with our intuition. It is human nature to see patterns where none exist, or to attribute cause to what is simply random fluctuation of naturally occurring events. Even if factors operate to affect outcomes in one direction or another, random variation is prone to exaggeration and it is human nature to interpret such exaggeration as though it has a cause. It is perfectly reasonable, for instance, to observe a football team have remarkable success one year, only to do disastrously the following year, and to ascribe to this all manner of explanations *apart from random variation* (i.e. regression to the mean)! The same can be said for science: if focus, or disproportionate attention, is given to a specific issue, unusual or extreme outcomes may be observed, but they need not have any sound causal origins and could well be a statistical quirk and product of the data generation processes.

Attention to some patterns may be skewed simply due to the data acquisition process. If data collected for one purpose are used for another, or data acquired in a certain way is more likely to feature specific structures due to the underlying data generation or data acquisition processes, then statistical artefact will arise. This may seriously distort any meaningful causal interpretation. With an exponential growth of information available and the commensurate big data revolution, without careful consideration and reflection upon the provenance of the plethora of data that will no doubt be subject to extensive 'fishing', we are likely to be provided with many misguided claims of 'robust' associations; despite the distinction between prediction (as primarily used for big data) and causal inference, it will be human nature to attribute erroneously cause and effect!

8. UNEXPLAINED RESIDUALS MODELS

Learning objectives

- Understand the implications of causal inference in seeking implicit conditioning in models
- Be aware of one sophisticated attempt to model longitudinal data that introduces problems

Explicit conditioning

Within multivariable regression, the inclusion of multiple variables explicitly involves the notion of conditioning because the estimated coefficient of any one variable in the model is conditional on the simultaneous consideration of all other variables in the same model.

We recall that, when we regress y on x whilst adjusting for z, we are asking: *What is the relationship of x with y whilst keeping z constant?* The assumption made is that the x-y relationship is the same for all values of z, i.e. the relationship is conditionally 'independent' of z. We need to think carefully about the implications of holding z constant. For instance, when $z = x^2$ (i.e. we have the quadratic model $y = \beta_0 + \beta_1 x + \beta_2 x^2$) we do not interpret the coefficient for x (β_1) as though x^2 were constant, but instead consider the *joint* interpretation of both coefficients for x and x^2. In such a scenario, β_1 and β_2 would be considered simultaneously when seeking to understand the x-y relationship.

This may be familiar and it is straightforward to interpret *curvilinear* relationships that can also be visualised, but not all implications of conditioning within multivariable regression are as trivial. We examine a situation in longitudinal data analysis where a sophisticated form of conditioning is used to address the problem of collinearity, which can arise in longitudinal data. However, this method introduces more problems than it resolves, since it fails to consider the causal framework in which it seeks to operate. This is an example of the pitfalls of sophisticated methodology used in the absence of any careful reflection on the role of causal inference.

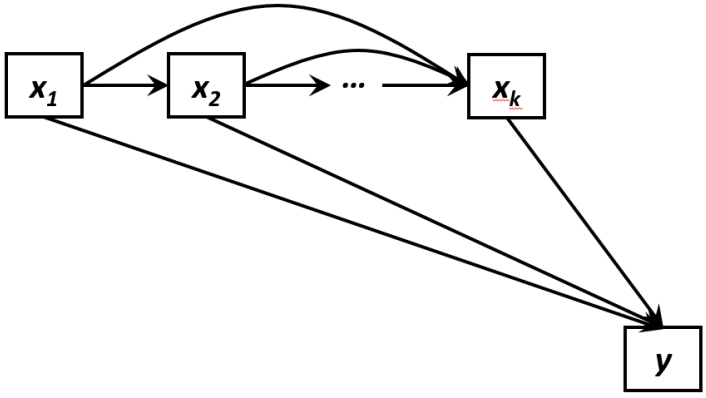
This example arises for what is termed **'unexplained residuals' models**, also known in parts of the epidemiology literature as **conditional models**, though this name is less helpful since, as noted previously, all multivariable models are explicitly 'conditional'.

'Unexplained residuals' (UR) models

UR models have been proposed as a way of evaluating the relationship between an exposure x measured longitudinally (e.g. x_1, x_2, … x_k, for k repeated measures) and a future outcome y (often referred to as a *distal* outcome); such a relationship is represented in Figure 1 as a DAG.

Accurately modelling such a scenario may help identify and quantify important periods of change or growth in x that affect the outcome y. Using multivariable regression, researchers would (ideally) treat each longitudinal measure as a separate exposure that is confounded by all prior exposure measures; the total number of models would thus be equal to the total number of time points at which the exposure is measured.

Figure 1: DAG of longitudinal exposure (i.e. x_1, x_2, … x_k, for k repeated measures) in relation to distal outcome y



As repeated measures are likely serially correlated, some models could potentially suffer from high levels of collinearity. We are mindful of (non-parametric) causal relationships amongst variables prior to analysis, but concern arises that (parametric) collinearity may be sufficient to impact adversely on multivariable regression model precision for those models containing two or more exposure variables.

UR models are proposed to address these concerns, as well as purportedly to quantify the total causal effect for *each* measurement of the exposure within a single model[79].

Explaining UR models

The simplest longitudinal scenario involves two exposures, x_1 and x_2, and a distal outcome, y. To estimate the total causal effect of *each* exposure variable on y, the following two standard regression models ($\hat{y}_s^{(i)}$, for $i = 1{,}2$) would typically be constructed:

$$\hat{y}_s^{(1)} = \hat{\alpha}_0^{(1)} + \hat{\boldsymbol{\alpha}}_{\boldsymbol{x1}}^{(\mathbf{1})}\boldsymbol{x_1} \qquad \text{Eq.1}$$

$$\hat{y}_s^{(2)} = \hat{\alpha}_0^{(2)} + \hat{\alpha}_{x1}^{(2)}x_1 + \hat{\boldsymbol{\alpha}}_{\boldsymbol{x2}}^{(\mathbf{2})}\boldsymbol{x_2} \qquad \text{Eq.2}$$

For each model, we interpret only the estimated coefficient of the *last* measurement as a **total causal effect**, as all previous values of the exposure would be mediated by later values, thereby potentially invoking bias due to the reversal paradox[23]. To bypass the need for several models, it has been suggested that the information contained within these two separate models may be captured in one overall regression model by using 'unexplained residuals'[79]. It is claimed that such a model allows the researcher to quantify the effects on the outcome of the initial exposure x_1 and subsequent *changes* in x within a single model.

The modelling process requires two relatively straightforward steps:

1. x_2 is regressed on x_1 (i.e. $x_2 = \hat{\gamma}_0^{(2)} + \hat{\gamma}_{x1}^{(2)}x_1 + e_{x2}$), which produces a measure of each observation's 'expected' value of x_2 as predicted by its value of x_1. The difference between expected and actual values of x_2 (i.e. $\hat{\gamma}_0^{(2)} + \hat{\gamma}_{x1}^{(2)}x_1$) amounts to the residual term e_{x2}.
2. y is regressed on the initial exposure x_1 and subsequent residual term e_{x2}:

$$\hat{y}_r^{(2)} = \hat{\lambda}_0^{(2)} + \hat{\boldsymbol{\lambda}}_{\boldsymbol{x1}}^{(\mathbf{2})}\boldsymbol{x_1} + \hat{\boldsymbol{\lambda}}_{\boldsymbol{ex2}}^{(\mathbf{2})}\boldsymbol{e_{x2}} \qquad \text{Eq.3}$$

The 'unexplained residuals' (UR) model (Eq.3) is meant to have the following advantages[79]:

- It produces the same **predicted outcomes** as the standard regression model in Eq.2 that includes both x_1 and x_2 (i.e. $\hat{y}_s^{(2)} = \hat{y}_r^{(2)}$);
- The estimated **model coefficient** values produced by individual standard regression models (Eq.1 and Eq.2) are equal to those estimated within the UR model (i.e. $\hat{\alpha}_{x1}^{(1)} = \hat{\lambda}_{x1}^{(2)}$ and $\hat{\alpha}_{x2}^{(2)} = \hat{\lambda}_{ex2}^{(2)}$), allowing for multiple coefficients to be interpreted as **total causal effects** within a single model;
- It provides insight (via the coefficient $\hat{\lambda}_{ex2}^{(2)}$) into the additional influence of x increasing *more than expected* upon y; and
- The initial exposure x_1 and residual increase e_2 are mathematically independent (i.e. orthogonal).

Succinctly, the two models $\hat{y}_S^{(2)}$ and $\hat{y}_{UR}^{(2)}$ are algebraically equivalent, but $\hat{y}_{UR}^{(2)}$ does not suffer collinearity and makes interpretation of the separate influence of the initial measurement of the exposure x (i.e. x_1) *and* subsequent changes in x more straightforward than do (multiple) standard regression models $\hat{y}_S^{(1)}$ and $\hat{y}_S^{(2)}$.The approach outlined here may be extended to any number of measurements of an exposure variable and the same properties will be upheld, further minimising the impacts of collinearity.

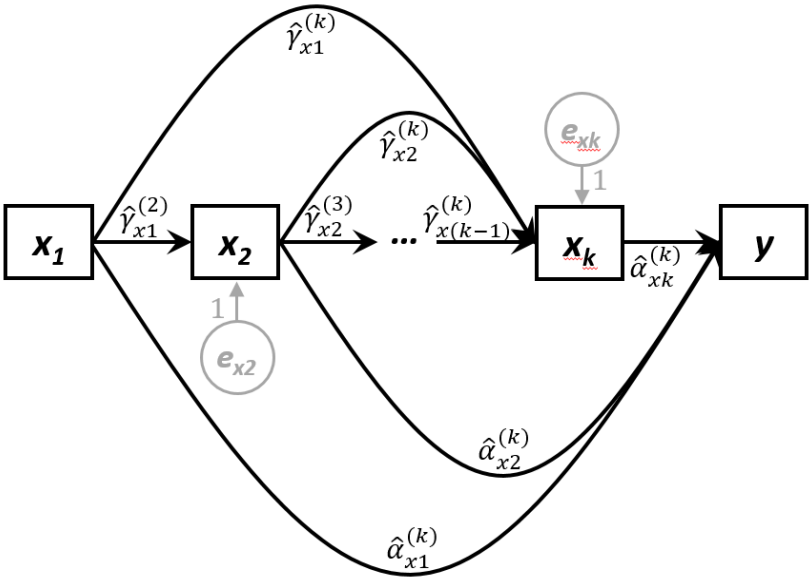
A causal framework

Within a causal framework, the unique properties of UR models are easy to visualise. In many respects, a UR model is akin to a structural equation model (SEM)[80], which is a (linearly) parametric DAG. In Figure 2, we order the nodes temporally and add the UR terms and appropriate regression coefficients (representing the estimated direct effects between pairs of variables) to our original DAG that includes no additional confounding variables (Figure 1). The coefficients amongst measurements of x are obtained via the regression of each measurement of x on all previous measurements $x_1, \ldots, x_{i-1}$, whereas the coefficients between measurements of x and the outcome y are obtained via the standard regression model which includes all measurements of x. In this way, each endogenous node on the graph (except y) in Figure 2 is represented as a linear combination of all preceding nodes and an error term.

From Figure 2, we can see why the UR modelling process works in the absence of additional confounding.

If we naively model $x_1, x_2, \ldots, x_k$ simultaneously, only the coefficient of the final measurement x_k could be interpreted as a total causal effect on y; the coefficients of $x_1, \ldots, x_{k-1}$ would represent only the direct effects of each measurement on y, as all future measurements would fully mediate the respective relationship. In graphical model language: all 'backdoor' paths[18] would be blocked by preceding measurements.

Figure 2: A (linear parametric) DAG depicting k longitudinal measurements of exposure x (x_1, x_2, … x_k, for k repeated measures), one distal outcome y, and a time-invariant confounder m, with regression coefficients and UR terms added.



By modelling $x_1, e_{x2}, \ldots e_{xk}$ (as in a UR model), we encounter no mediation problems since, by construction, the UR terms remain wholly independent of the other terms in the model. In fact, by placing the UR model in a causal framework, we can see that the UR terms $e_{x2}, \ldots e_{xk}$ are essentially instrumental variables (IVs)[81] for $x_2, \ldots x_k$, respectively, produced by the modelling process.

Take x_2 as an example, where $k = 3$. The total effect of x_2 on y encompasses the direct effect from $x_2 \rightarrow y$ and all indirect effects (of which there is only one in this scenario): $x_2 \rightarrow x_3 \rightarrow y$. Table 1 gives the total effects of x_2 on y and of e_{x2} on y (calculated via the method of path coefficients[82]), with both total effects decomposed into their respective direct and indirect effects. From Table 1, we

see that the total effect of x_2 on y is equal to the total effect of e_{x2} on y; this is because there are no direct paths between e_{x2} and y, and all indirect paths pass through x_2 (with the coefficient of e_{x2} on x_2 being equal to one).

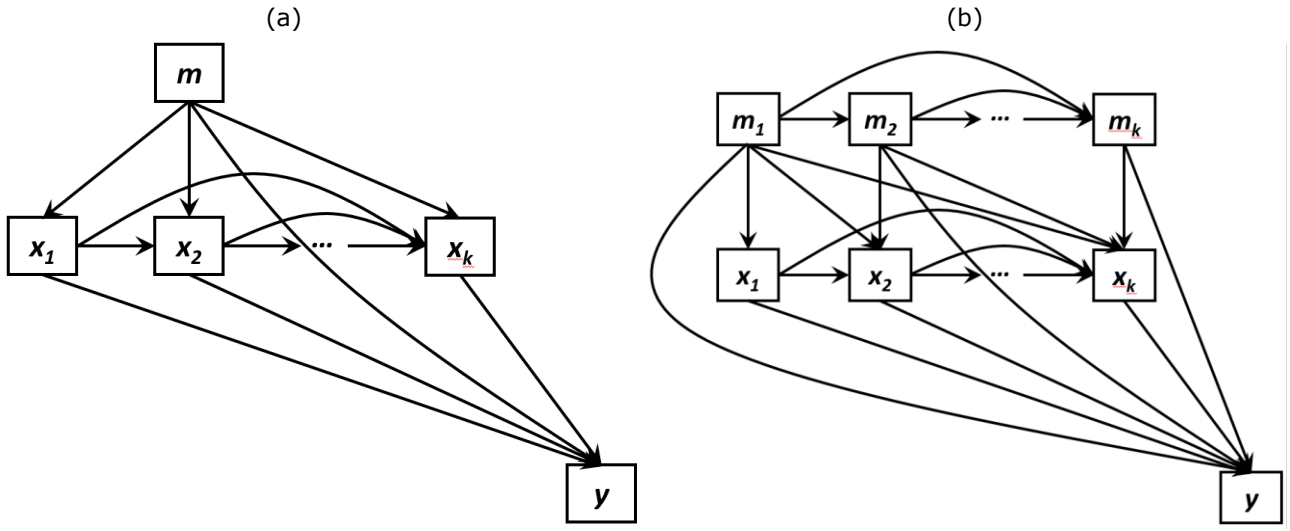
Exposure:	Path:		Effect size:	Total effect:
x_2				
	Direct:	$x_2 \rightarrow y$	$\hat{\alpha}_{x2}^{(3)}$	$\hat{\alpha}_{x2}^{(3)} + \hat{\gamma}_{x2}^{(3)} \cdot \hat{\alpha}_{x3}^{(3)}$
	Indirect:	$x_2 \rightarrow x_3 \rightarrow y$	$\hat{\gamma}_{x2}^{(3)} \cdot \hat{\alpha}_{x3}^{(3)}$	
e_{x2}				
	Direct:	n/a		
	Indirect:	$e_{x2} \rightarrow x_2 \rightarrow y$	$1 \cdot \hat{\alpha}_{x2}^{(3)}$	$\hat{\alpha}_{x2}^{(3)} + \hat{\gamma}_{x2}^{(3)} \cdot \hat{\alpha}_{x3}^{(3)}$
		$e_{x2} \rightarrow x_2 \rightarrow x_3 \rightarrow y$	$1 \cdot \hat{\gamma}_{x2}^{(3)} \cdot \hat{\alpha}_{x3}^{(3)}$	

Table 1: Total effect of x_2 on y estimated by a standard regression model compared to total effect of e_{x2} on y estimated by an equivalent UR model (Figure 1b, with $k = 3$).

Additional confounders

Researchers have extended the original UR model by adjusting for additional confounders (other than prior measurements of the exposure x), but until recently[83;84] there has been no thorough discussion or analysis of this issue. Additional confounding (over and above all prior exposures) may be either **time-*invariant*** (Figure 3a) or **time-*variant*** (Figure 3b); incorrect adjustment for either can lead to biased causal inferences[84].

Figure 3: DAG of longitudinal exposure (i.e. x_1, x_2, … x_k, for k repeated measures) in relation to distal outcome y, with: (a) one time-invariant confounder m; and (b) k time-variant confounders $m_1, m_2, \ldots, m_k$.



UR models only produce equivalent coefficients to those of standard regression models when a time-invariant confounder is adjusted for during steps 1 and 2 in the model-creation process; if the time-invariant confounder is adjusted for only in step 1 (i.e. when generating each UR term e_{xi}) or only in step 2 (i.e. in the overall UR model $\hat{y}_k^{(2)}$), model estimates may be *inferentially biased* (though not *statistically biased*).

Additionally, UR models can accommodate a time-variant confounder, although the process is more intensive. UR terms must be created for the confounder itself, with each term adjusted for all previous values of the time-variant confounder *and* the exposure; UR terms must also be created

for the exposure, adjusted for all previous values of the exposure and all previous and current values of the time-variant confounder. The outcome is modelled as a function of the initial value of the exposure and all its subsequent UR terms, and the initial value of the confounder and all its subsequent UR terms. As with a time-invariant confounder, if these (substantial) adjustments are not made, model estimates may suffer *inferential bias*, resulting in incorrect causal claims.

Whereas the individual standard regression models are functions of highly correlated, causally linked longitudinal exposures, the composite UR model is instead a function of mathematically independent 'competing exposures'.

Interpretability issues

Whilst UR models reduce collinearity, this is found generally to be quite modest as most longitudinal data may be relatively sparse across large time intervals and therefore not highly collinear. Further, despite claims to the contrary, these models offer no additional insight into periods of change in an exposure in relation to a distal outcome[85]. Perhaps most importantly, the explicit conditioning of each UR term on all previous terms renders independent interpretation of coefficients impossible and leads to a nonsensical situation in which variables in a UR model are interpreted as simultaneously increasing and being held constant.

More philosophically, terms in a UR model are independent of one another as an artefact of ordinary least-squares regression, though this is unlikely to be an accurate representation of real-world exposure variables. Many of these, such as body size, exhibit a consistent, cumulative presence that is only manifest at the discrete time points at which it is measured; these measurements are thus distinct only because of the discretisation of time within the measurement processes adopted. Moreover, in auxological studies, the phenomenon of so-called compensatory (or 'catch up') growth has been well documented, with accelerated growth being observed in individuals who begin with a low value of some measure, e.g. birthweight. Therefore, although convenient and mathematically sound, it may be unrealistic to model a longitudinal exposure in a way that implies complete independence between its initial value and all its subsequent changes. Moreover, the process of creating UR models leads to artificially reduced standard errors, which may mislead researchers about the true precision of the estimated total effect sizes.

Summary

Focusing on the perceived 'problems' with collinearity, without paying sufficient attention to the causal framework in which we are operating, only distract from staying clear-headed about the robust application of multivariable regression models that yield meaningful causal interpretation. It remains imperative amidst any form of statistical wizardry that we are anchored to our notions of what constitutes robust causal inference. 'Unexplained residuals' models may seem to overcome collinearity, providing mathematically reduced standard errors, but this is both misleading (smoke and mirrors) and can lead to the loss of any meaning if not applied carefully. There is no actual gain in causal insight obtained from UR models, and merely presenting a series of separate model estimates in a single model runs the risk of misinterpretation.

9. MATHEMATICAL COUPLING: ANALYSIS OF CHANGE WITH RESPECT TO BASELINE

Learning objectives

- Understand mathematical coupling (MC) in the analysis of change with respect to initial value
- Know strategies to overcome MC in the analysis of change with respect to initial value

Mathematical Coupling

In its simplest form, **mathematical coupling** (MC) is the phenomenon where the *null hypothesis* is distorted due to an **algebraic relationship** between two or more variables that are analysed by correlation or regression. Due to this distortion, any test of the null hypothesis (i.e. that the regression coefficient is zero) will be biased[86], as will any corresponding inferences[87-89]. Hypothesis testing becomes invalid because coupled variables are no longer mathematically independent.

MC most noticeably occurs when a new variable is constructed from a mathematical transformation of another, e.g. through addition, subtraction, multiplication or division[86;87;89-93]. Examples include change variables (e.g. change between baseline and follow-up) and ratio variables, either where one variable is divided by another (e.g. prevalence proportions) or divided by a function of another (e.g. body mass index [BMI], where weight in kilograms is divided by height in meters squared). MC then arises if these constructed variables are analysed with respect to any of their component variables using correlation or regression (e.g. comparing two prevalence rates, which share the same denominator, or predicting BMI from height). The effects of MC are known and have been stated in a range of clinical domains[94;95], yet its consequences remain frequently overlooked. We examine the issue of MC for ratio variables later, but for now, we look at the context where one is interested in the relation between change and initial value.

The relation between change and initial value

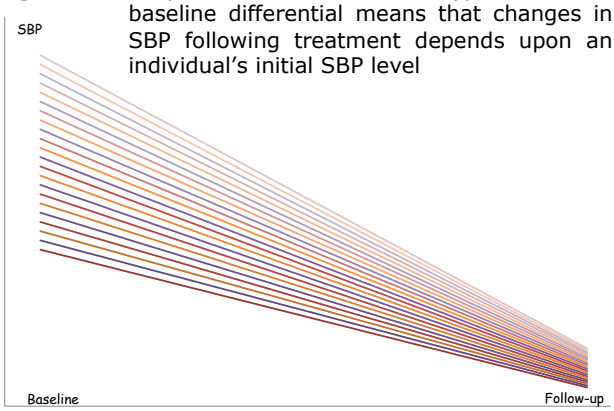
The most widely recognised illustration of MC arises in the analysis of change with respect to initial value. The relation between initial disease status and change following an intervention has attracted considerable interest in clinical research. What seems a relatively simple issue is deceptively complex, and the obvious strategies for analysing such data are highly problematic[96].

For instance, we ask: Do individuals with higher initial systolic blood pressure (SBP) experience greater SBP reduction following intervention?

We are therefore asking if there is a **differential effect**, where changes in SBP depend on patients' initial SBP level (Figure 1).

One could test the relation between *change* and *initial value* using correlation or regression, yet this would be inappropriate for the reasons previously explained.

Figure 1: Response to treatment for hypertension: a baseline differential means that changes in SBP following treatment depends upon an individual's initial SBP level



Despite many articles and medical statistics textbooks warning against correlating or regressing change on initial value, many researchers (including many statisticians) still overlook the problem and/or are not aware how MC can cause bias. For instance, a decrease in the oxygenation index has been 'shown' to be proportional to baseline oxygenation index in infants with persistent

pulmonary artery hypertension undergoing nitric oxide therapy[97]; changes in plasma concentration of plasminogen-activator inhibitor type 1 (PAI-1) are correlated with PAI-1 concentration before treatment in postmenopausal women treated with oral oestrogen therapy[98]; change in CD4(+)Ki67(+) T-cells is significantly correlated with the change in total CD4(+) T-cells in HIV-infected subjects undergoing antiretroviral therapy[99]; percentage change in bone mineral density in the spine is highly correlated with baseline values in women receiving either hormone replacement therapy or (perhaps this should have been a giveaway?) placebo[100]; the percentage of patients with 3-month Barthel scores ≥95 are highly correlated with the percentage with small artery disease and the percentage using tobacco[101]; and there is a strong, negative correlation between height and BMI[102].

At first glance, it is far from clear what the problem is, which is perhaps why so many researchers continue to make the same analytical mistakes. Many of those who recognise there are problems with analysing change with respect to baseline (statisticians included), often mistakenly attribute the problem entirely to *regression to the mean* (RTM); the issue of mathematical or causal coupling is then overlooked completely. We have previously examined the impacts of RTM and note that it is not only RTM that gives rise to the problems outlined here.

Explaining the impact of MC for the relation between change and initial value

In a seminal article in 1962, Oldham warned against testing the effect of anti-hypertensive therapy with respect to baseline blood pressure[103]. One of his arguments, subsequently repeated many times since[86;104], is that for two independent random numbers (e.g. z and y) with identical standard deviations, there will be a strong correlation (averaging $1/\sqrt{2} \approx 0.71$) between their difference ($z - y$) and either variable (positive if correlated with z; negative if correlated with y). The regular assumption that the null is zero is entirely untrue. Any estimated relationship between change and baseline or follow-up will comprise an element of artefact plus an element of true effect (if non-zero); since the artefact is sizeable, it will likely dominate the true effect. Oldham set out to explain this as follows.

Let x_b be pre-treatment (baseline) values and x_f post-treatment (follow-up) values. The Pearson correlation between change ($x_b - x_f$) and pre-treatment value (x_b) is given by[103]:

$$Corr[x_b - x_f, x_b] = \rho_{(b-f)f} = \frac{s_b - \rho_{bf} s_f}{\sqrt{s_b^2 + s_f^2 - 2\rho_{bf} s_b s_f}} \quad \text{Eq.1}$$

where s_b^2 is the variance of the x_b, s_f^2 is the variance of x_f, and ρ_{bf} is the correlation between baseline (x_b) and follow-up (x_f). If s_b^2 and s_f^2 are equal, Eq.1 reduces to:

$$\rho_{(b-f)f} = \frac{1-\rho_{bf}}{\sqrt{2}}. \quad \text{Eq.2}$$

Eq.2 shows that unless ρ_{bf} is unity, $\rho_{(b-f)f}$ **will never be zero**; when $\rho_{bf} < 1$ (which is highly likely for repeated measurements on the same individuals) the correlation between baseline and change will always be positive. In fact, if $\rho_{bf} \approx 0$, i.e. there is very poor correlation between pre- and post-treatment values (or we have random numbers with equal standard deviation), the positive association between baseline and change will be large $(1/\sqrt{2} \approx 0.71)$.

In biomedicine, since both x_b and x_f are always subject to measurement error and/or biological variation, $\rho_{bf} < 1$ and $\rho_{(b-f)f} > 0$. Under H_0 (where change is <u>not</u> related to initial value) the correlation of change with initial value will never be zero. The same is true for the regression of change on initial value – the model coefficient is never zero. This effect is striking, and for even modest sample sizes it will be statistically significant; however, this will likely be nothing but artefact.

<u>Oldham's method</u>

Oldham suggested that testing the hypothesis that treatment is associated with baseline should be carried out by plotting change against the mean of the pre- and post-test values, and <u>not</u> against the baseline values. For instance, if pre-treatment (baseline) SBP is denoted as x_b and post-treatment (follow-up) SBP as x_f, SBP reduction following an anti-hypertensive medication will be $x_b - x_f$, and mean SBP will be $(x_b + x_f)/2$. To address whether greater baseline SBP is related to a greater reduction in SBP following treatment, Oldham's method tests the correlation between $x_b - x_f$ and $(x_b + x_f)/2$ instead of testing the correlation between $x_b - x_f$ and x_b. The Pearson correlation between **change** and **average** is[103]:

$$Corr\left[x_b - x_f, (x_b + x_f)/2\right] = \frac{s_b^2 - s_f^2}{\sqrt{\left(s_b^2 + s_f^2\right)^2 - 4\rho_{bf}^2 s_b^2 s_f^2}} \qquad \text{Eq.3}$$

where s_b^2 is the variance of x_b, s_f^2 is the variance of x_f, and ρ_{bf} is the correlation between baseline (x_b) and follow-up (x_f).

The numerator in Eq.3 indicates that Oldham's method is a test of the **difference in the variances** between the repeated measurements, where the two variances may also be correlated (captured in the denominator). If there is no difference in the variances of pre-treatment SBP (x_b) and post-treatment SBP (x_f), the correlation from Oldham's method will be zero, i.e. the treatment effect (SBP change) is not associated with the mean SBP value (and hence not associated with baseline or follow-up values, as explained below).

The rationale behind Oldham's method is that if, on average, a greater reduction in SBP is observed for greater values of SBP at baseline, then post-treatment SBP values will become 'closer' together, i.e. the post-treatment variance (s_f^2) will *shrink* and be smaller than the pre-treatment variance (s_b^2). Conversely, if there is a **differential effect** (either due to a *differential effect of treatment*, or a *differential physiological response to treatment*, i.e. greater or smaller effects in those with greater or lesser disease severity, respectively), this will manifest as a **change of variances** between the first and second measure. If there is no difference in the variances before and after treatment, there is little or no evidence of a differential effect.

MC remains between the change, $x_b - x_f$, and the mean, $(x_b + x_f)/2$, because each expression contains terms in common with the other (x_b and $\pm x_f$), but this specific approach nullifies the impact of MC. To understand why a test of the relation between change and mean yields a correct null hypothesis, one must appreciate that the sum of any two variables with equal variances is always uncorrelated to the difference between them. This can be shown using vector geometry[105], though such insights are left for extra-curricular reading!

Oldham's strategy has been proposed previously, as early as 1939 by Morgan and Pitman[106;107], though specifically in the context of testing the equivalence of two variances. Much later, in 1985, the same approach also became the basis of the Bland and Altman approach for comparing two methods of measurement[108], though it may not be immediately recognised as a solution to MC.

Note: The problem has been addressed by the hypothesis being re-framed in terms of testing the difference in variances, rather than simply correlating or regressing change with baseline.

A multilevel solution to the relationship between change and initial value

MC is removed completely if one models the repeated measures (baseline and follow-up) using multilevel modelling[109;110]. This approach also allows for the inclusion of additional covariates and more than two measurement occasions (i.e. studies with multiple follow-up occasions). Critically, however, **time must be centred** (see below).

Considering the blood pressure scenario for illustration, initial and post-treatment SBP measures are at the lower level of the multilevel model and individuals are at the upper level. The model covariate *Time* represents the initial and post-treatment measurement occasions, and the correlation between the variance of the random intercept and the variance of the random slope for the covariate *Time* indicates the relation between baseline disease status (intercept) and treatment effect (slope). The model is:

$$\mathrm{SBP}_{ij} = \beta_{0ij} + \beta_{1j} Time_{ij} \qquad \text{Eq.5}$$

Different parameterisations of *Time* yield different results, and model misspecification will lead to the same consequences as with MC[111]. For instance, when *Time* is coded as 0 (initial) and 1 (post-treatment), the correlation between the random intercept and random slope is equivalent to the correlation between change and baseline ($\rho_{f-b,b}$ adopting earlier notation), since the random slope variance is estimated from differences between baseline and follow-up, $f - b$, whilst the random intercept variance is estimated at baseline, b. If *Time* is coded in reverse, i.e. -1 (initial) and 0 (post-treatment), the correlation between random intercept and random slope is equivalent to the correlation between change and follow-up, for similar reasoning. Only if *Time* is centred, i.e. -/+ 0.5 (initial / post-treatment), is the correlation between random intercept and random slope equivalent to Oldham's method ($\rho_{f-b,(b+f)/2}$), since the change variance is estimated as before, but the intercept variance is now estimated at the point midway between baseline and follow-up, i.e. at their mean value of $(f + b)/2$.

It is common for analysts to employ multilevel models but fail to centre their time variable; this matters if the covariance for the random intercept and slope is to be interpreted!

Summary

MC is ubiquitous, yet despite impacting research in many instances, its consequences are poorly recognised. MC for the analysis of change with respect to initial value can be overcome by using multilevel models, provided there is careful consideration of model parametrisation.

10. MATHEMATICAL COUPLING: ANALYSIS OF RATIO VARIABLES

Learning objectives

- Understand how mathematical coupling (MC) may occur for ratio variables
- Know strategies that *might* overcome MC for ratio index variables with common denominators

Mathematical Coupling

As stated previously, **mathematical coupling** (MC) is the phenomenon where the *null hypothesis* is distorted due to an **algebraic relationship** between two or more variables that are analysed by correlation or regression. Most noticeably, MC occurs when a new variable is constructed by the mathematical transformation of others, including multiplication or division[86;87;89-93]. The most ubiquitous example that involves multiplication / division is body mass index (BMI), where weight in kilograms is divided by height in meters squared. MC arises if constructed variables are analysed with respect to any of their component variables using correlation or regression (e.g. examining the relationship between BMI and height). We examine the motivation for the use of constructed ratio variables, and initially highlight the pitfalls this can generate due to MC. Later, we consider the many broader causal implications of **composite variable confounding**.

Ratio index variables

A *ratio index variable* is a new variable derived from the division of one variable by another. In epidemiology, one is often concerned with prevalence and incidence (counts of total cases per population and counts of new cases per population per unit time, respectively), which are ratios that capture the *relative* extent of a condition (e.g. prevalence of obesity, incidence of mortality) by accounting for differences in population sizes. In medicine, many variables are generated as ratios to capture human features (e.g. obesity), acknowledging that humans vary due to genetic predisposition (e.g. height). Hence, such ratios seek to capture a *relative* construct (e.g. BMI as a measure of weight relative to height-squared). The concept of what is *relative* in both contexts is seeking to standardise a measure with respect to a perceived 'norm', such as average body height or a typical cross-section of society.

The potential for MC when constructing and evaluating ratio variables by correlation or regression is huge, but is largely overlooked. The implications of MC amongst ratio variables are numerous and far reaching, yet almost no attention is given to the artefacts generated within epidemiology, or observational research more generally.

Explaining the impact of MC for ratio variables

To illustrate, consider three random variables (x, y and z) that are uncorrelated with each other and have identical standard deviations. It can be shown that the correlation of x/z with $y/z \approx 0.5$[112]. Put simply, a strong correlation will exist between two variables when divided by the same denominator, even if they otherwise have nothing in common. As with the analysis of change, the assumption that the null is zero for the correlation or regression of ratio variables that share a common denominator is entirely false. Any estimated relationship will comprise an element of true effect (if non-zero) plus artefact; the latter will again be sizeable and likely dominate.

Some of the most important outcomes and exposures in epidemiology are ratios, which is why the impacts of MC are so crucial. For instance, MC will occur via the common denominator of *population*

at risk when investigating the relationship between incidence rates of two or more diseases (e.g. cancer and diabetes), or when investigating the relationship between the incidence of a disease and the prevalence of an exposure for that disease (e.g. asthma and the proportion of overcrowded households). For modest sample sizes, a highly significant association can be observed, even if the relationship is entirely artefact.

Potential solutions to MC due to common denominators

This problem first came to light over a hundred years ago when Pearson warned would-be-analysts to be wary of a 'spurious correlation' that arises between two ratio variables with a common denominator[112]. The term mathematical coupling did not appear in Pearson's paper, as the term was not coined until many years later[90].

To tackle the problem, Pearson suggested that analysts should calculate the *partial correlation* between numerators (disease counts) whilst 'adjusting' (within a regression model) for the common denominator (population counts), rather than analysing the two ratios directly[67]. Poisson regression automatically advocates this approach by encouraging analysts to model counts with a denominator 'offset' included as a model covariate (logged to match the Poisson log-link). Consequently, Poisson multivariable regression avoids the adverse impacts of MC, though this is merely fortuitous and not by intentional design.

Following Pearson's warning, Neyman reiterated that ratio variable numerators and denominators should be separated and analysed as Pearson suggested, seeking partial correlations[68]. In 1947, Fisher set out to '*illustrate the extreme simplicity*' of dealing with '*problems concerned with the relation of a part to the whole*'[69;113], advocating the same solution as Pearson and Neyman, though his paper serves to illustrate only the complexity of such problems[69;113].

The importance of a causal framework

Fisher used data that contained the body and heart weights of cats from a group of digitalis assays. Since *body-weight* comprises *heart-weight*, their ratio is compositional. Fisher adopted *body weight* as the dependent variable and *heart weight* as the independent variable. This may seem reasonable from a physiological viewpoint if the heart is thought to be the driver of circulation and its size therefore determines capacity for growth, driving total body size. However, if body size determines the volume of blood required, this would determine the required size of the heart to service circulation. From this perspective, *body weight* would be the ancestor to *heart weight,* requiring the implied regression model to be the opposite to that proposed by Fisher. Since one variable comprises the other, their causal relationship is impossible to resolve unequivocally.

Thinking causally presents a bigger problem, however, when one considers the role of *sex*. It is reasonable to assume that *sex* determines both *heart weight* and *body weight* because genes that determine *sex* are likely to influence body development in a way as to influence both *heart weight* and *body weight*. Asking if there are sex differences in the relationship between *heart weight* and *body weight*, as Fisher did, then *sex* is the exposure of interest, regardless of how we view the causal relationship between *heart weight* and *body weight*: within a regression model, one weight will be the outcome and the other is a mediator. In any causal framework considering *heart weight*, *body weight* and *sex*, it will always be inappropriate in a multivariable model seeking to examine the causal effect of *sex* to include *heart weight* as a covariate if *body weight* is the outcome, or

vice versa. Fisher's original question is therefore intractable within a multivariable model due to the conflict the model generates within a causal framework – something that escaped this giant of statistics because a formal understanding of causal inference, as described through graphical model theory, had not been developed back then.

Other ratio variable constructs

MC arises amongst constructed ratio variables if each possess common components as numerator or denominator, i.e. x/z is also coupled with z/y and z/x is also coupled with z/y. Coupling will similarly occur if either numerator or denominator is a function of common elements. For instance, w/h^2 is coupled to any expression of w or h. If $w = \mathbf{weight}$ and $h = \mathbf{height}$, $w/h^2 = \mathbf{body\ mass\ index}$.

The proposed solution when there are **common denominators** is to decouple the denominator, and 'adjust' for it directly within a multivariable regression analysis to obtain the *partial correlation*. Other examples of MC will need different solutions, though how to proceed may not be immediately apparent.

Summary

MC is ubiquitous, yet despite impacting research in many instances, its consequences are poorly recognised. MC amongst variables that are ratio generated with common denominators can be approached differently, whereby the ratio variable components are separated and the common denominator is treated as a separate covariate within a multivariable regression model. However, this only proves informative if the resulting regression model makes sense in a causal framework, i.e. the common denominator is a confounder of the exposure of interest, and not a mediator, as was the problem in Fisher's paper advocating the separation of ratio variables into their components. Rubin recently argued that Fisher's advice might not be wise for every instance, since the model assumptions can often be violated[114;115]. If we have doubts and concerns with statistical 'giants' such as Fisher, then these problems are clearly not trivial, which perhaps explains why confusion and controversies persist.

Although widely employed in biomedical research, constructed ratio variables are problematic and present many challenges. This points to the concern that any composite variable, i.e. a variable that is constructed through addition, subtraction, multiplication or division of other variables, is potentially problematic when seeking to place the composite variable within a causal framework. This leads on nicely to the issues of **composite variable confounding**.

11. COMPOSITE VARIABLE CONFOUNDING

Learning objectives

- Understand how composite variable confounding (CVC) arises
- Recognise implicit CVC from the construction of composite variables
- Understand how to address CVC for the analysis of change with respect to baseline exposures

Composite variable confounding

Composite variable confounding (CVC) is a form of bias that results from the naïve analysis and interpretation of composite variables (such as change variables, ratio variables, and other constructed variables) without separately considering the causal influences and consequences of each individual component. Mathematical coupling can therefore be considered a special case of CVC, where the composite variable has been analysed in relation to one of its own components, or a function thereof (See Figure 1: panel A + B). But any variable that has been constructed from two or more other variables (whether exposure or outcome) is prone to create problems from CVC. To avoid this, each constituent component should be considered individually within a causal framework. Even where no bias due to explicit mathematical coupling is introduced, the conflation of different causal relationships into a single summary measure that is then analysed within a multivariable regression model can create substantial interpretational challenges (as we will see in the example below).

When causal inference is sought, the causal structure of a dataset should be postulated *a priori*, which we can depict using a DAG. DAG-data consistency can be evaluated prior to statistical evaluation[13] and for each exposure-outcome relationship of interest, we can identify sets (formally known as the 'minimally sufficient adjustment sets') of variables that control for confounding. The regression coefficient of the 'exposure' variable of interest may then be interpreted as an estimate of the **total causal effect** (notwithstanding the problems of measurement error, missing data, and residual confounding, etc.). Although it is increasingly recognised that robust causal estimates require use of a robust causal framework, such practices remain uncommon. CVC is therefore likely to arise often, yet will be poorly recognised, and there are likely many important instances that have not yet been uncovered.

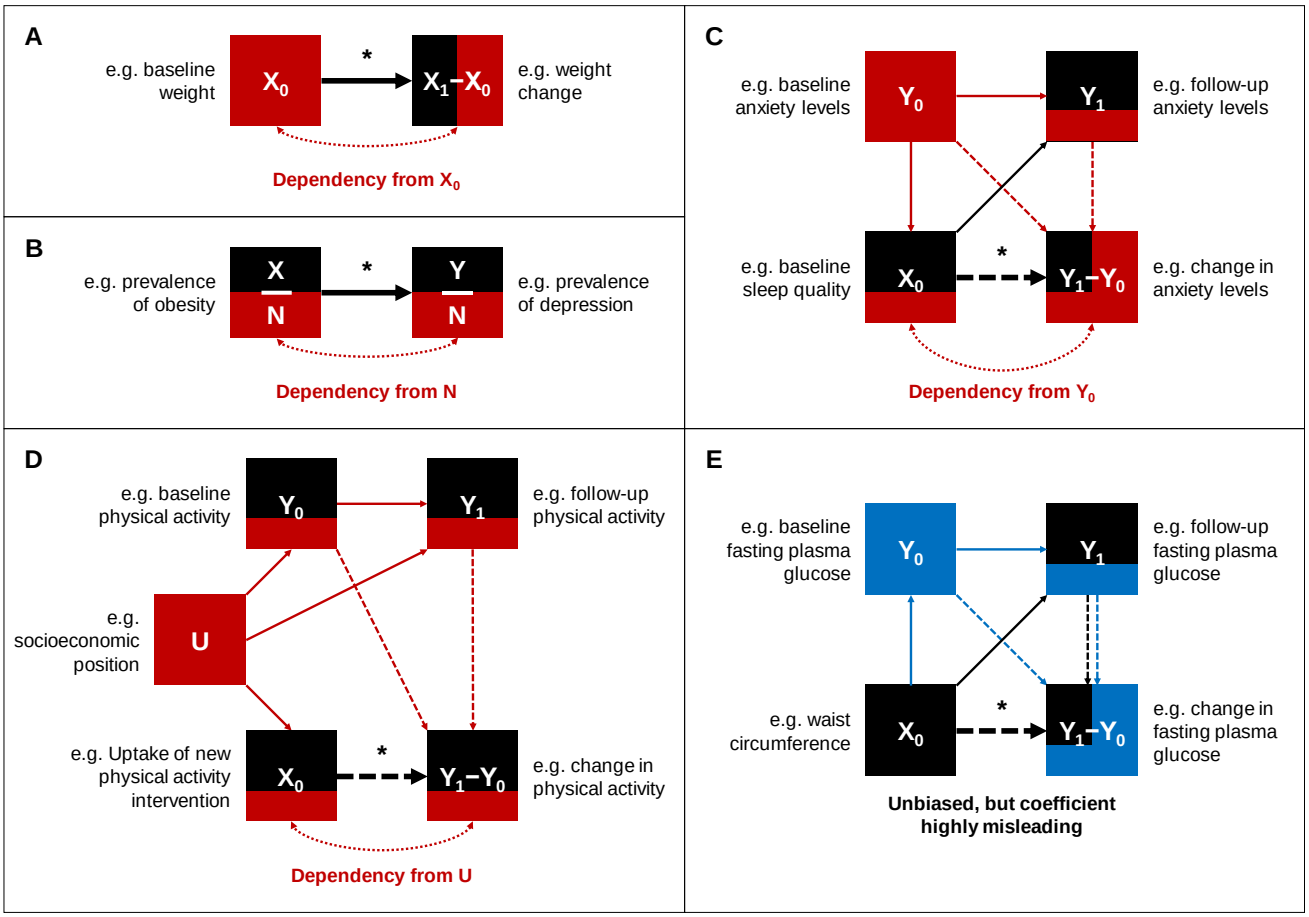
Illustration for the construction of a change variable

To illustrate CVC, we examine the evaluation of a composite *change* variable, though the principle extends to many other instances of constructed variables, including ratio variables such as BMI.

A key issue within the analysis of change is that the *change variable* is treated as a single concept, when in fact 'change' comprises information about both baseline and follow-up; each should therefore be considered separately within a causal framework. Formally, the change 'effect' is captured entirely by the follow-up outcome alone. This is immediately apparent when considering the context of an RCT. Since there is no relationship between the baseline 'exposure' (the intervention) and the baseline outcome (values of which have been randomised between the treatment arms), it is entirely sufficient to examine the relationship between the intervention and the follow-up outcome. For observational data, a relationship between the baseline exposure and outcome is however very likely. The creation of a change variable, by subtracting follow-up from baseline, is an attempt to resolve this unwanted correlation and obtain a relative or 'standardised'

measure of change. This approach was developed long before the development of modern causal inference methods; use of which now make the potential issues rather obvious.

Figure 1: Schematic examples of composite variable confounding. The causal relationship under test is marked with an asterisk sign (*). Panels A and B depict two examples of mathematical coupling. In **A**, the focal relationship (X_1-X_0 ~ X_0) is biased by a dependency between the composite outcome (X_1-X_0; e.g. weight change) and the exposure (X_0, e.g. baseline weight) resulting from the outcome having been algebraically constructed from the exposure. In **B**, the focal relationship (Y/N ~ X/N) is biased by an explicit dependency between the composite outcome (Y/N, e.g. prevalence of depression) and the composite exposure (X/N, e.g. prevalence of obesity), resulting from both having been algebraically constructed from a shared denominator variable (N, e.g. regional population). Panels C, D, and E depict how composite variable confounding can occur even without the explicit problems of mathematical coupling. In **C**, the focal relationship (Y_1-Y_0 ~ X_0) is biased by a dependency between the composite outcome (Y_1-Y_0, e.g. change in anxiety levels) and the exposure (X_0, e.g. baseline sleep quality) resulting from the exposure and the follow-up outcome (Y_1, e.g. follow-up anxiety levels) being mutually determined by the baseline outcome (Y_0, e.g. baseline anxiety levels). In **D**, the focal relationship (Y_1-Y_0 ~ X_0) is biased by a dependency between the composite outcome (Y_1-Y_0, e.g. change in physical activity) and the exposure (X_0, e.g. uptake of a new physical activity intervention) resulting from the exposure and the follow-up outcome (Y_1, e.g. follow-up physical activity) being mutually determined by an unobserved confounder (U; e.g. socioeconomic position). In **E**, the focal relationship (Y_1-Y_0 ~ X_0) is unbiased, but the coefficient is highly misleading. It neither represents the total causal effect of the baseline exposure (X_0, e.g. baseline waist circumference) on the follow-up outcome (Y_1, e.g. follow-up serum insulin concentration) nor the causal effect of the baseline exposure on the follow-up outcome conditional on baseline outcome (Y_0, e.g. baseline serum insulin concentration). The regression coefficient instead represents a competing sum of the two causal paths (shown in blue and black), which may commonly be negative; it is not clear when this estimate would be anything other than misleading!



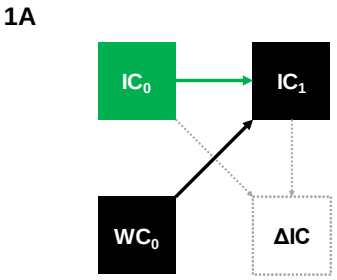
An alternative way to analyse change that separates the baseline and follow-up information is to model follow-up while adjusting for baseline (an approach that is often called an 'ANCOVA'). We consider the example of evaluating change in serum insulin concentration in relation to baseline

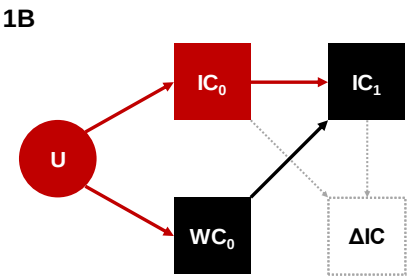
waist circumference. We focus less on the clinical relevance of this example and more on the insights it offers about CVC and the benefits of thinking within a causal framework.

A typical dataset would contain the exposure variable (waist circumference) at baseline (e.g. WC_0), the outcome (serum insulin concentration) at baseline and follow-up (e.g. IC_0 and IC_1), and a derived *change* variable (e.g. $\Delta IC = IC_1 - IC_0$). For this example, we will explore and interpret the coefficient produced from a linear model of change in outcome regressed on baseline exposure ($\Delta IC \sim WC_0$) for a series of scenarios. This would typically be interpreted (explicitly, or more often implicitly) as the **total causal effect** of baseline waist circumference on the *change* in insulin concentration.

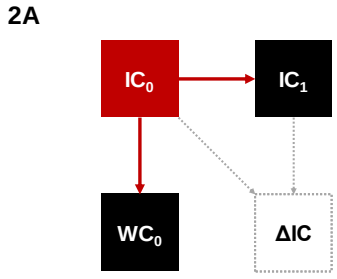
Figure 2: Causal scenarios for the relationship of change in insulin concentration ($\Delta IC = IC_1 - IC_0$) regressed on baseline waist circumference WC_0. In part 1, IC_0 is a competing exposure for the effect of WC_0 on IC_1. (1A) WC_0 causes follow-up insulin IC_1 but not baseline insulin IC_0; (1B) WC_0 causes follow-up insulin IC_1 but not baseline insulin IC_0; WC_0 and IC_0 are caused by one or more unobserved (i.e. latent) factors, collectively denoted U. In part 2, IC_0 is a confounder for the effect of WC_0 on IC_1. (2A) IC_0 causes both WC_0 and IC_1; WC_0 does not cause IC_1. (2B) IC_0 causes both WC_0 and IC_1 and WC_0 causes IC_1. (2C) IC_0 causes both WC_0 and IC_1; WC_0 does not cause IC_1; WC_0 and IC_0 are caused by one or more unobserved (i.e. latent) factors (U). (2D) IC_0 causes both WC_0 and IC_1; WC_0 causes IC_1; WC_0 and IC_0 are caused by one or more unobserved (i.e. latent) factors (U). In part 3, IC_0 is a mediator for the effect of WC_0 on IC_1. (3A) WC_0 causes IC_0 and IC_0 causes IC_1 but WC_0 does not cause IC_1. (3B) WC_0 causes both IC_0 and IC1, and IC_0 causes IC_1. (3C) WC_0 causes IC_0 and IC_0 causes IC_1; WC_0 does not cause IC_1; WC_0 and IC_0 are caused by one or more unobserved (i.e. latent) factors (U). (3D) WC_0 causes both IC_0 and IC1, and IC_0 causes IC_1; WC_0 and IC_0 are caused by one or more unobserved (i.e. latent) factors (U). In all scenarios, change in insulin (ΔIC) is explained entirely by IC_0 and/or IC_1, and is not caused by any other variables, hence the arcs to ΔIC, and the composite variable itself, are shown in grey.

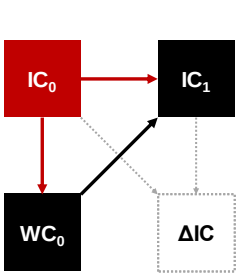
IC_0 as **competing exposure**:

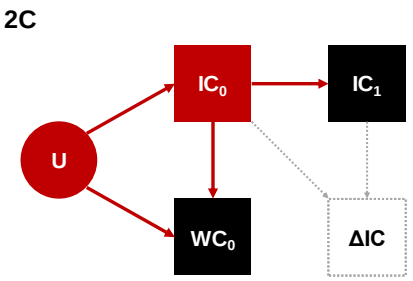


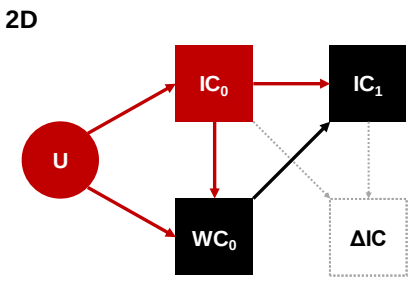


IC_0 as **confounder**:

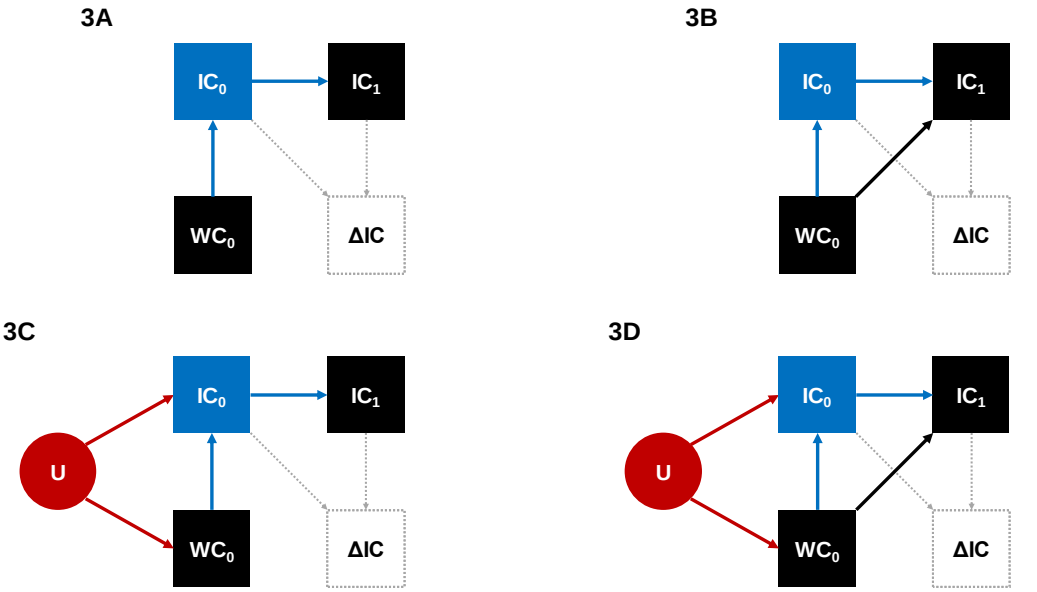








IC_0 as **mediator**:



To explore the various implications, we consider several different causal scenarios depicted by the ten DAGs in Figure 2. In each scenario, we consider the implications of interpreting the regression coefficient of the change in outcome regressed on baseline exposure as the **total causal effect** (as would be standard, even if subliminal).

***Interpretation of β_1 as total causal effect in the model:* $\Delta IC = \beta_0 + \beta_1 WC_0 + \ldots$**

Scenario 1A: The coefficient would be correctly interpreted as the *total causal effect* of WC_0 on ΔIC. The absence of an arc between WC_0 and IC_0 means the estimated relationship with ΔIC would only include the activity on IC_1. Note, this scenario essentially depicts a randomised controlled trial, and is what we are seeking to emulate for observational analyses of change. Fortunately for those analysing RCT data, the lack of association between the intervention and the outcome at baseline, means they can analyse a composite change outcome without risk of CVC, because the '$-IC_0$' component contributes nothing to the association.

Scenario 1B: The estimated causal effect of WC_0 on IC_1 is confounded by one or more unobserved variables (U) that influence IC_1 via their effects on IC_0. The estimated causal relationship of WC_0 on the composite outcome ΔIC will be similarly confounded by the latent variables and cannot therefore be interpreted as *total causal effect*. This risk of confounding arises for observational studies due to the lack of randomisation.

Scenario 2A: The estimated effect of WC_0 on IC_1 is confounded by the causal effect of IC_0 on WC_0 and IC_1 and cannot therefore be interpreted as the *total causal effect*. In truth, there is no causal effect of WC_0 on IC_1, so the estimated effect for the composite outcome ΔIC will be entirely due to confounding.

Scenario 2B: As with Scenario 2A, the estimated effect of WC_0 on IC_1 is confounded by the causal effect of IC_0 on WC_0 and IC_1 and cannot therefore be interpreted as the *total causal effect*. On this occasion, however, there is a causal effect of WC_0 and IC_1. Alas, by analysing the composite outcome ΔIC, this will be conflated with the confounded effects of IC_0 on WC_0 and IC_1.

Scenario 2C: As with Scenarios 2A and 2B, the estimated effect of WC_0 on IC_1 is confounded by the causal effect of IC_0 on WC_0 and IC_1, but there is now additional confounding by one or more unobserved variables (U) that influence IC_1 via their effects on IC_0. Since there is no causal effect of WC_0 on IC_1, the estimated effect for the composite outcome ΔIC will be entirely due to confounding, and cannot be interpreted as the *total causal effect*.

Scenario 2D: As with Scenario 2C, the estimated effect of WC_0 on IC_1 is confounded by the causal effect of IC_0 on WC_0 and IC_1 and one or more unobserved variables (U) that influence IC_1 via their effects on IC_0. On this occasion, however, there is a causal effect of WC_0 and IC_1. Alas, by analysing the composite outcome ΔIC, this will be conflated with the confounded effects of IC_0 on WC_0 and IC_1 and of the confounded effects of U on WC_0 and IC_1 through IC_0, so the coefficient cannot be interpreted as the *total causal effect.*

Scenario 3A: This scenario is unaffected by confounding but the coefficient is highly misleading. The causal effect of WC_0 on IC_1 is entirely mediated through IC_0. Alas, by analysing the composite outcome ΔIC, this effect must compete mathematically with the causal effect of WC_0 on IC_0 so the coefficient cannot be interpreted as the *total causal effect*. Without a separate effect of WC_0 on IC_0, the coefficient here will always be *negative*, since the diluted effect on IC_1 acting through IC_0 will be smaller than the full effect acting directly on IC_0.

Scenario 3B: As with Scenario 3B, this scenario is unaffected by confounding but presents a high risk of severe interpretational bias. The causal effect of WC_0 on IC_1 is partly mediated through IC_0. Alas, by analysing the composite outcome ΔIC, this effect – and the unmediated effect of WC_0 on IC_1 – must compete mathematically with the causal effect of WC_0 on IC_0. The coefficient cannot therefore be interpreted as the *total causal effect*. Depending on the relative sizes of the mediated and unmediated effects, the coefficient in this situation may often be *negative*.

Scenario 3C: As with Scenario 3A, the causal effect of WC_0 on IC_1 is entirely mediated through IC_0, but there is also confounding by one or more unobserved variables (U) that influence IC_1 via their effects on IC_0. The coefficient cannot therefore be interpreted as the *total causal effect*. By analysing the composite outcome ΔIC, the true causal effect must compete mathematically with the causal effect of WC_0 on IC_0 and the confounded association between WC_0 and IC_0 due to U. Depending on the relative sizes of these effects, the coefficient in this situation may often be *negative*.

Scenario 3D: As with Scenario 3B, the causal effect of WC_0 on IC_1 is partly mediated through IC_0, but there is also confounding by one or more unobserved variables (U) that influence IC_1 via their effects on IC_0. The coefficient cannot therefore be interpreted as the *total causal effect*. By analysing the composite outcome ΔIC the causal effects of WC_0 on IC_1 (both mediated and unmediated through IC_0) must compete mathematically with the causal effect of WC_0 on IC_0 and the confounded association between WC_0 and IC_0 due to U. Depending on the relative sizes of these effects, the coefficient in this situation may often be *negative*.

Alternative analytical strategies: interpretation of in the model: $IC_1 = \beta_0 + \beta_1 WC_0 + \ldots$

If the analytical focus is shifted from the composite outcome ΔIC to the follow-up outcome IC_1, the problems of CVC can be avoided. Considering the underlying causal framework in each scenario can then provide insight on how best to analyse the effect of WC_0 on IC_1.

In Scenario 1B, IC_0 is a proxy confounder for the unobserved confounder(s) U of the relationship between WC_0 and IC_1. The *total causal effect* of WC_0 on IC_1 could therefore be estimated by adjusting for IC_0 as a proxy confounder (i.e. $IC_1 \sim WC_0 + IC_0$). Residual confounding is likely, however, as we are not adjusting for the true confounders directly.

In Scenarios 2A and 2B, IC_0 confounds the relationship between WC_0 and IC_1. The *total causal effect* of WC_0 on IC_1 could therefore be estimated by adjusting for IC_0 (i.e. $IC_1 \sim WC_0 + IC_0$).

In Scenario 2C and 2D, IC_0 is both a proxy for the unobserved confounder(s) U, and a genuine confounder of the relationship between WC_0 and IC_1. The *total causal effect* of WC_0 on IC_1 could therefore be estimated by adjusting for IC_0 as both genuine and proxy confounder (i.e. $IC_1 \sim WC_0 + IC_0$). Residual confounding is likely, however, as we have not adjusted for the true confounders U directly.

In Scenarios 3A and 3B, IC_0 is a mediator of the relationship between WC_0 and IC_1. The *total causal effect* of WC_0 on IC_1 could therefore be estimated in the univariate model without adjustment for IC_0 (i.e. $IC_1 \sim WC_0$)

In Scenarios 3C and 3D, IC_0 is both a proxy for the unobserved confounder(s) U and a mediator of the relationship between WC_0 and IC_1. Here, there is no robust analytical means for obtaining the unconfounded *total causal effect* of baseline WC_0 on IC_1. Without further adjustment, the estimate is confounded by U, but adjusting for IC_0 as a proxy confounder would be erroneous, and would risk invoking the reversal paradox[23]. To obtain a robust estimate, we would have to obtain additional information on U, either by collecting further data or deriving estimates from the literature (and performing simulations).

Summary

CVC is ubiquitous, yet despite impacting research in many guises, its consequences and even its existence are hardly recognised. In the example, the standard analytical strategies are highly problematic, and more thoughtful approaches are needed that place the separate components of composite variables within a causal framework. To illustrate CVC, we examined the *analysis of change*, but there are many other instances where a composite variable is analysed as a single concept with insufficient consideration given to the unique causal relationships of the composite variable's constituent components. There are likely many undiscovered scenarios affected by CVC and solutions may not always be apparent.

In our 'toy' example, we did not dwell on the specific clinical context, but made a general point of how CVC arises. However, in most genuine clinical situations where a composite measure of *change* is evaluated in relation to baseline exposures, there are few conceivable observational instances where the baseline exposure is totally unrelated to the baseline outcome. If the exposure is believed to cause the outcome at follow-up, it seems most likely that the baseline exposure would also cause the outcome at baseline (i.e. Scenario 3B in Figure 2). That said, reality may often be closer to that of Scenario 3D (Figure 2), where there is unmeasured confounding. The latent confounding may also have a direct causal impact on the follow-up outcome, making the true effect even more difficult to observe. In any event, there seems no obvious alternative analytical strategy for regressing follow-up outcome on baseline exposure to ensure we obtain a robust estimate of the *total causal effect* of baseline exposure on *change*, as depicted by the follow-up outcome

measure. In general, it is often impossible to obtain the *total causal effect* without carefully considering and measuring all the relevant confounding variables.

A limited appreciation of CVC means we encounter many analyses that yield meaningless and/or misleading findings. It is therefore vital to be vigilant and committed to robust practices of causal inference. Problems with CVC - including mathematical coupling - are potentially amongst the most ubiquitous and severe methodological errors in biomedical research.

13. SPARSE OUTCOMES & MIXTURE MODELLING

Learning objectives

- Be aware of statistical challenges in modelling sparse count data (i.e. an 'excess' of zeros)
- Understand the importance of data generation to guide model parameterisation and selection

Modelling count data with 'excess' zeros

In a variety of research domains, where the outcome is counts, it is common to find an 'excess' of zeros relative to standard count distributions; this occurs regularly in epidemiology. Ridout *et al*. have reviewed several methods to address excess zeros, particularly in relation to the Poisson distribution[116]. The **zero-inflated Poisson** (ZiP) model[117-119] is one such strategy, where the overall distribution is a mixture of two distributions: one with a central location (i.e. mean) of zero (i.e. a 'spike' of zeros) and the other with a non-zero central location to be estimated empirically (i.e. a regular Poisson that may depend upon covariates)[120]. The proportion of each distribution are determined empirically and may be thought of as separate models. The **zero-inflated binomial** (ZiB) model is another strategy[121;122], akin to the ZiP model, but with a bounded number of counts. The ZiP and ZiB models have been used extensively in many research domains and are amongst the most often considered, though there have been recent developments with generic mixture models, also known as **latent class models** or **discrete latent variable models**.

A generic **mixture model** determines several *latent classes* or subgroups of data, the optimum choice of which is typically informed by log-likelihood statistics. Model parameters of each class, along with their contribution to the combined outcome distribution, are determined empirically. ZiP/ZiB models are limited forms of mixture models: a mix of exactly two distributions where one comprises entirely zeros. Mixture models extend beyond the zero-inflated models to allow any number of distributions, where no one distribution is constrained to be identically zero (i.e. there is no 'spike' of zeros, unless the model empirically determines one distribution to have a mean of zero, which amounts to the same thing).

In examining mixture models, we consider the following (frequently overlooked) statistical issues that have relevance to all complex modelling strategies:

- The choice of **outcome distribution** is crucial.
- **Over-dispersion** should always be considered.
- **Predicted outcomes** can have greater clinical relevance than likelihood statistics.
- **Covariates** in the **distribution** model must be considered in the **class membership** model.

Adopting a Poisson distribution for all count outcomes is naïve, as the Binomial distribution may be better; over-dispersion can be a consequence of clustering that arises implicitly, even if the data are not obviously hierarchical or clustered; assessing model-fit by likelihood statistics overlooks clinical context in judging a model; and omitting covariates from the class membership model that are considered in the distribution model imposes constraints that may lead to biased models. Moreover, both likelihood statistics and predicted outcomes may not indicate an 'ideal' model, since different model parameterisations can yield near-identical fit statistics and near-

identical predicted outcomes; model selection must then be aided by knowledge of the **data generation process**.

Dental Example Dataset

In dental research, an established indicator of a person's oral health status involves counting the number of decayed (d/D), missing (m/M), and filled (f/F) deciduous 'milk teeth' (t) or permanent teeth (T), yielding the measure of *dmft* or *DMFT*[123]. The *dmft* count may range between 0 and 20, whereas the *DMFT* count may range between 0 and 32. Amongst healthy individuals, or during the early stages of dentition development, there is potential for an excess number of zero *dmft/DMFT* counts. For illustration, we consider a prospective study in the Brazilian urban area of Belo Horizonte during the early 1990's, which examined different dental caries prevention methods amongst 797 school children aged 7 years at the start of the study[123]. Data were recorded for the eight deciduous molars only, so *dmft* counts ranged between 0 and 8. The research focus was how different intervention methods prevent caries incidence (new lesions). Interventions comprised: (1) oral health education; (2) enrichment of the school diet with rice bran; (3) mouthwash with 0.2% sodium fluoride (NaF) solution; (4) oral hygiene; (5) all the interventions combined; or (6) none of the interventions (control). The proposed outcome was *change* in the *dmft* count from baseline.

There are limitations to this study because school allocation, although random, involved only one school per intervention arm, which is insufficient for adequate cluster-randomisation; thus, baseline differences in mean *dmft* across intervention groups may not have been due to chance. The authors purportedly sought to accommodate baseline mean differences in disease levels amongst schools by using ANCOVA, though you should recall that this only accommodates *within-group* heterogeneity whilst assuming *between-group* baseline mean differences are minimal due to randomisation. As too few schools were randomised, observed baseline mean outcome differences amongst groups were found, which could yield biased results (i.e. Lord's paradox[124-126]). As the original study findings are questionable, these data are only considered for illustrative purposes. Skrondal and Rabe-Hesketh analysed the follow-up data only[127] to illustrate the use of generic mixture models. We examine the same data to explore these methods too, though we do not seek to draw any meaningful inferences.

Statistical considerations of model parameterisation

Choice of distribution

Böhning *et al*.[123] used their data to argue that ZiP models are useful in evaluating intervention effects on dental caries when data exhibit an excess of zero counts. However, Skrondal and Rabe-Hesketh questioned the use of the Poisson distribution in this instance, as the outcome adopted represents the number of *dmft* ('successes') out of a total of eight deciduous molars ('trials')[127]. The ZiB model was compared with the ZiP model, which revealed that the latter typically predicted unrealistically long tails and the former performed better. The binomial outcome is preferred in this instance, as the count index is bounded at eight.

Over-dispersion

Over-dispersion (i.e. where the outcome distribution has a heavier tail than expected) is a common issue in surveys where units are cluster-sampled (e.g. children nested within schools, as in the

example data). Many situations arise where count data form an implicit hierarchy or clustering, even if not intentional or by design. For instance, in conducting a survey, each field worker forms a cluster; this may give rise to over-dispersion that is overlooked. When the Brazilian dental data were examined following the models proposed by Böhning *et al*. (i.e. ZiP) and by Skrondal and Rabe-Hesketh (i.e. ZiB), along with over-dispersed equivalents, the latter consistently performed better as per likelihood-based model-fit criteria.

Choice of model-fit criteria: beyond likelihood statistics

The likelihood statistics often considered when determining model fit are the Bayesian Information Criterion (BIC) and Akaike's Information Criterion (AIC), both of which incorporate a measure of model parsimony to provide a trade-off between model complexity and how well the model fits the data[128]. These likelihood-based model-fit criteria are recommended, though criteria based on the difference between observed and predicted outcomes should also be considered, with relevance to clinically relevant thresholds along the outcome scale. For instance, the transition from zero to one represents *onset* of disease in longitudinal data and increased *prevalence* of disease in cross-sectional data. The tails of a distribution indicate disease *progression* for longitudinal data and disease *extent* for cross-sectional data. The crossing of any 'critical' threshold might distinguish between 'high' or 'low' risk groups for targeted intervention. A threshold may represent a point of no return (e.g. mortality or tooth loss in the dental example). Generally, **model fit assessment should have clinical relevance** and ought to be more than evaluation of log-likelihood statistics.

Class prediction in zero-inflated models

Extending standard ZiP/ZiB models to include class prediction by covariates involves replacing the parameter for the two-distribution proportions (depicting the extent of belonging either to the zero-bin or to the standard distribution) with a function of the available covariates, just as the standard distribution is a function of covariates[120]. The class membership model is a logistic regression model with covariates. Standard or over-dispersed distributions may apply.

An enormously overlooked issue is that ZiP/ZiB models are in fact problematic if covariates in the distribution (non-zero) part of the model are not considered as class predictors (i.e. to determine whether individuals belong to the zero-bin or the distribution part of the model). There is no explicit discussion of this in the literature until that by Gilthorpe *et al*.[120]. This is fundamental, since the proportion of zero counts (i.e. the proportion of disease-free children in the example dataset) is otherwise constrained and erroneous models may arise, leading to inappropriate interpretations of the data. To illustrate, data were simulated (similar in nature to those observed in dental caries studies) to reveal the extent of bias that results if covariates that are deemed necessary in the distribution part of ZiP/ZiB models are not also considered as class predictors.

Consider the covariate *sex* and two-stage data simulation where *dmft* outcomes comprise 50,000 boys and 50,000 girls: 20% of the boys have a *dmft* of zero, with the rest taking values from a Poisson distribution with mean 2; 80% of the girls have a *dmft* of zero, with the rest taking values from a Poisson distribution with mean 1. Extending the model with *sex* predicting class membership is therefore essential, though typically overlooked. We examine how unreliable the standard ZiP model is for this scenario. Results are presented in Table 1.

The number of girls in the zero-bin is poorly predicted by the ZiP model (22.87%, far from the simulated true of 80%). The distribution mean for girls is also far from true (estimated as 0.27, opposed to the true value of 1.00). For girls, the incorrectly specified ZiP model yields considerable deviation from truth in terms of size, shape and central location of the distribution, yet overall predicted counts are indistinguishable from those simulated.

Table 1: Model-fit criteria for the ZiP model undertaken with the simulated data

	Simulated True	ZiP Estimated
Log-Likelihood	-100,088 †	-111,700.74
BIC	200,212 †	223,436.02
AIC	200,184 †	223,407.48
Observed – Predicted zero counts	0	1,573.74
Girls		
Proportion in the Zero-bin	80%	22.87%
Distribution mean *dmft* count (95% CI)	1	0.27 (0.26, 0.28)
Boys		
Proportion in the Zero-bin	20%	22.87%
Distribution mean *dmft* count (95% CI)	2	2.03 (1.97, 2.07)

ZiP – standard zero-inflated Poisson model with *sex* as a covariate in the non-zero part only (not as a class predictor); BIC – Bayesian Information Criterion; AIC – Akaike's Information Criterion; †true log-likelihood, BIC and AIC are based on the asymptotic likelihood, which was maximised numerically.

Different parameterisations and inferences are feasible whilst predicted counts hardly differ with no sizeable difference in likelihood-based model-fit criteria; it is thus difficult, if not impossible, to decide upon an 'ideal' model using model-fit criteria alone. This issue is not limited to zero-inflated models.

Generic mixture models

Zero-inflated models are a special case of the generic mixture model. The most general form of a mixture model is where each class adopts the standard distribution (i.e. not constrained to be zero) and class membership is potentially informed by covariates and becomes a multinomial logistic regression model[128]. Many model options are available, though not all are interpretable; in some instances, models may not be identifiable. For instance, if a covariate impacts differently within each latent class, and if class membership is predicted by this covariate, model interpretation is challenging (even if the model is identifiable) because circularity arises regarding the conditionality of the relationship of covariate parameters in the distribution parts *and* the class membership part of the same model.

Distinguishing between different model parameterisations

We illustrate the problems that can arise when seeking to distinguish between different model parameterisations of both zero-inflated and generic mixture models by re-evaluating the Brazilian dataset[123]. We consider a range of standard binomial, zero-inflated, and generic mixture models (for the complete set of models see Gilthorpe *et al*.[120]). We consider binomial models since the outcome is bounded above, and we allow for over-dispersion since the study data are inherently clustered (children within schools). Table 2 summarises observed and predicted counts for the best model from each of the **standard binomial**, **zero-inflated** and **generic mixture** models (the best of each type was always over-dispersed).

Table 2: Binomial regression models: observed and predicted counts of *dmft* along with model fit criteria assessments

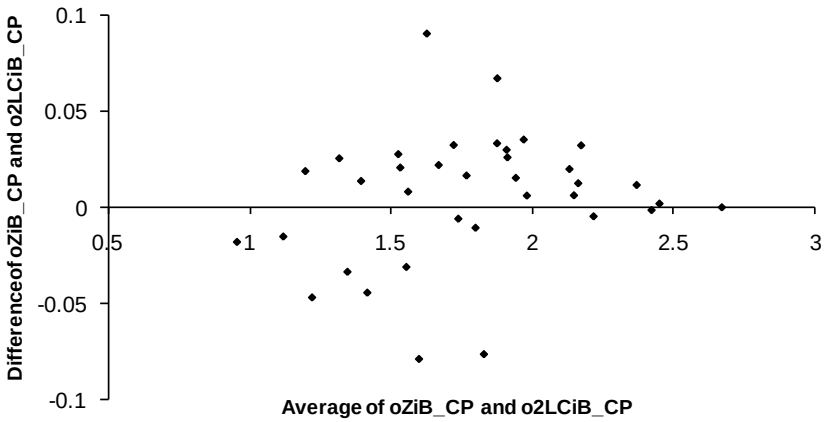
Dmft	**N**	**oB**	**oZiB_CP**	**o2LCiB_CP**
0	*231*	217.29	230.45	227.35
1	*163*	189.08	156.15	169.70
2	*140*	146.75	151.17	137.68
3	*116*	104.64	117.24	114.26
4	*70*	68.40	76.08	80.15
5	*55*	40.22	41.21	43.99
6	*22*	20.41	17.95	18.05
7	-	8.18	5.73	5.07
8	-	2.02	1.03	0.75
Total	*797*	797.00	797.00	797.00
Class size		-	†15.36%	‡44.62%
Log-Likelihood		-1,402.61	-1,393.78	-1,386.48
BIC		2,872.03	2,914.49	2,906.57
AIC		2,825.22	2,825.56	2,812.96

oB: over-dispersed binomial model; oZiB_CP: over-dispersed zero-inflated binomial model with the same covariates in the non-zero part and predicting class membership; o2LCiB_CP: over-dispersed 2-class mixture model with class independent covariates predicting class membership; †size of the zero-bin for zero-inflated models; ‡size of the 2nd latent class.

The BIC and AIC model-fit criteria do not agree as to the *best* model; they agree on the *worst* – the zero-inflated model – which is the best for the predicted number of zeros. Outcome-specific model-fit criteria do not generally agree with likelihood-based model-fit criteria, which highlights the important role that of clinically relevant model-fit assessment criteria have.

Nevertheless, given the disparities amongst all model-fit criteria, it seems difficult to choose an 'ideal' model. There are few differences in predicted counts, demonstrated by contrasting the two models with reasonable predicted outcomes (oZiB_CP and o2LCiB_CP): expected counts (predicted probabilities for all 36 types of children, i.e. 6 interventions × 2 genders × 3 ethnicities) are close (ρ=0.98), and a Bland-Altman plot[129] reveals no systematic bias (Figure 2).

Figure 2: Bland-Altman plot of contrast between oZiB_CP and o2LCiB_CP



Thus, selecting a 'preferred' model is less than obvious using likelihood-based model-fit criteria *or* predicted outcomes. Each model parameterisation has potentially different interpretation and blindly settling upon one could give rise to misleading inferences of the data. To inform model selection, we turn to *a priori* knowledge of the data generation process.

Data generation informs model parameterisation / selection

We ask whether zero-inflated models and generic mixture models might reflect different underlying processes generating the data. Model selection could then seek to distinguish between modelling strategies according to *a priori* hypotheses of data generation. This requires context-specific appreciation of the data being modelled. We consider how dental caries occur, i.e. how the dataset was generated, and which model is more plausible clinically.

Clinical context

For biomedical data in general, and caries data specifically, we consider the distinct roles of disease *onset* and *progression* in relation to observed data distributions, and look at how this might inform model choice. For instance, caries disease onset requires one tooth to become decayed, filled, or extracted (i.e. a *dmft* increment from 0 to 1). Thereafter, an increment to this score requires a second tooth to suffer a similar fate. It is known that some teeth and some tooth surfaces are more prone than others to the effects of the cariogenic environment (i.e. the level of oral hygiene maintained: amount and frequency of starch/sugar-rich snacking). For instance, first molars are more prone to caries than second molars; upper teeth more prone than lower teeth; pit and fissure surfaces are more prone than approximal or smooth surfaces[130].

The nature of the cariogenic exposure is also important, since different teeth have different caries risk depending on their morphology and position in the mouth relative to the salivary gland ducts and accessibility for tooth brushing. Moreover, teeth erupt or are shed (exfoliated) at different times, and the 'risk set' thus varies over time (i.e. the period 'at risk' may vary from one tooth to the next). Amongst adults, teeth may also be extracted for reasons that have little to do with caries (orthodontics), thereby initiating the diseased state for reasons unrelated to subsequent caries. **Caries *onset* and *progression* might therefore have different underlying risks**[131], for which there is substantial support in the dental research literature[131-133]. Selecting between zero-inflated and generic mixture models is informed by *a priori* knowledge of the causal processes underlying caries data generation.

Hypothetical underlying data generating scenarios

One hypothesis is that the cariogenic environment of the *individual* does not depend on whether a tooth has already been affected, hence it is reasonable to assume that underlying latent risks of disease *onset* and subsequent *progression* are identical. Whilst differences occur across individuals, generic mixture models are then suitable to describe 'subtypes' of individuals.

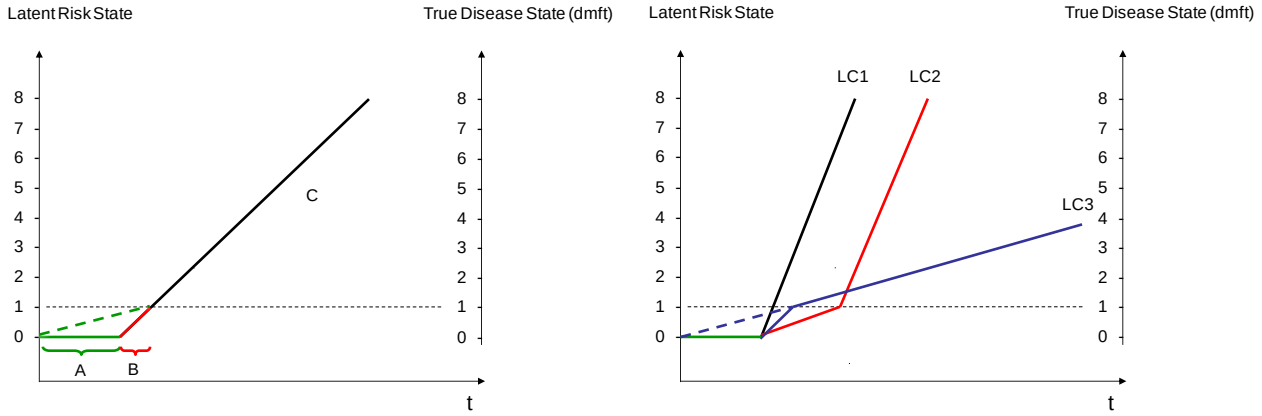
Another hypothesis is that underlying latent risk of disease differs across teeth or tooth surfaces, and there is a dual process of risk for disease *onset* and *progression*. A mixture of two outcome distributions would be manifest, where one has a central location of zero, and a ZiB model would be suitable to describe caries patterns.

Where underlying complexity warrants it, both hypotheses and modelling strategies may be valid and one adopts a generic mixture model with each latent class subdivided into a zero-bin and standard distribution.

To see how zero-inflated models and generic mixture models capture different mechanisms of data generation, consider Figure 1.

Figure 1: Hypothetical risk models for the onset and progression of *dmft*

Figure 1(i) Figure 1(ii)



Gradients represent the strength of underlying risks for disease *onset* and *progression*; A: period with no underlying risk of disease; B: period where disease-free individuals are susceptible to disease *onset*; C: period where individuals with existing disease are susceptible to disease *progression*; LC1: latent class one, sub-group of individuals with high risk of disease *onset* and *progression*; LC2: latent class two, sub-group of individuals with low risk of disease *onset* and high risk of disease *progression*; LC3: latent class three, sub-group of individuals with medium risk of disease *onset* and low risk of disease *progression*.

Figure 1(i) represents the situation where: (A) initially there is no latent risk (e.g. prior to any teeth erupting); (B) individuals experience the risk of disease *onset*, i.e. on course to yielding a non-zero *dmft/DMFT* score, though initially have a zero score; and (C) individuals with disease experience the same underlying latent risk of disease *progression* as for disease *onset*. Since there is a period where some teeth are not at risk of disease, the estimated underlying risk of disease *onset* (the dotted line) appears different to that for the risk of disease *progression*, even though the 'true' underlying latent risks are identical for the 'at risk' period.

Figure 1(ii) represents the situation where there are three latent sub-types of individuals, each with varying latent risks of disease *onset* and disease *progression*. For latent class one (LC1), the latent risk of disease *onset* and *progression* are identical. For latent class two (LC2), the underlying risk of disease *onset* is less than that of disease *progression*. The third latent class (LC3) exhibits the opposite, in that the underlying risk of disease *onset* is greater than that of subsequent disease *progression*. LC1 and LC2 exhibit near identical underlying latent risks of disease *progression* despite having different underlying risks of disease *onset*. When the period 'not at risk' is included, the estimated underlying latent risks of disease *onset* and *progression* appear to differ for LC1, whilst they appear similar for LC3 – both contrary to 'true' and entirely due to the 'not at risk' period being misclassified or misinterpreted.

External information: balance of evidence

Given the overwhelming evidence in the dental research literature that risks of caries *onset* and *progression* differ, it is the most appropriate strategy to adopt zero-inflated models. Consequently, the 'preferred' model for the Brazilian dataset is the over-dispersed zero-inflated binomial model with the same covariates in the non-zero model part predicting class membership (oZiB_CP).

This was the **least favoured model as per the likelihood-based model-fit criteria**, but most favoured as per the number of predicted zero counts, highlighting the synergy between clinical model-fit criteria and data generation informing model selection. This also reveals how misguided it might be to favour likelihood-based model-fit criteria in selecting 'preferred' models. Generally, it is misguided not to introduce **contextual knowledge** and allow this to drive model selection.

Discussion

Böhning *et al*. rightly argued that one needs to consider carefully the problem of excess zeros in dental data. A Poisson distribution is not ideal if counts represent the number of successes (*dmft*) out of a finite number of trials. Binomial outcomes are preferable for bounded count data. Further, where data are inherently clustered (even if not by intent), over-dispersion ought to be considered.

For zero-inflated models with class membership not also predicted by covariates in the distribution part of the model, there is potential for bias due to unintended implicit constraints. Adopting context-specific model-fit criteria for predicted outcomes has clinical relevance and chimes with the underlying data generation process. However, there may be no discernible model differences in terms of either likelihood-based model-fit criteria *or* predicted outcomes between certain zero-inflated and generic mixture models. The challenge is how to select an 'ideal' model. In general, *a priori* knowledge of data generation helps inform model parameterisation and model selection to yield meaningful model inference.

The issues outlined here for count data can occur for other outcome distributions. In general, model building and selection is not only a matter of model fit, but also an issue of contextualisation that requires *a prior* appreciation of the data generation processes.

14. LONGITUDINAL EXPOSURES & LATENT VARIABLE MODELLING

Learning objectives

- Appreciate the challenges in modelling time-varying longitudinal exposures
- Be aware of latent growth curve models (LGCMs) and growth mixture models (GMMs)
- Respecting the data generation process in modelling random structure

Longitudinal exposures

Longitudinal patterns of clinical or anthropological attributes are often explored in epidemiology to identify how early-life experiences might influence later-life morbidity or mortality: this is lifecourse research. Methodological challenges arise since what is oftentimes a longitudinal outcome becomes a time-varying exposure; the outcome is now a measure downstream of the series of exposures (the exposure may be recorded right up to and including the time of the outcome). This specific exposure-outcome framework creates challenges within standard regression regarding **causal inference** and **model estimation**.

Causal inference is challenging as there is uncertainty in how to interpret a longitudinal exposure; we question if we are focusing on 'critical periods' (for targeted intervention), 'accumulated impact' (overall dose-response, with interest in cumulative exposure), 'trajectories' (sequenced or ordered combinations of events that impact differently if experienced at different stages of life and/or in different time order and/or in combination with other experiences), or other complex features of longitudinal exposure with causal implications. There are recent developments to address some of these questions[134;135] that avoid the methodological flaw of conditioning on the outcome *prior* to seeking to interpret 'trajectory' plots (we described this earlier as invoking RTM). Model estimation is challenging due to issues of nonlinearity and the potential for homoscedasticity (i.e. non-constant error structure), combined with the many ways we might elicit 'features' of the data, as per the various causal inference questions just highlighted. There are pros and cons to the methods to model longitudinal exposures, with no one proving ideal for all circumstances.

Multilevel modelling (MLM) is a common method of estimation for longitudinal measures in health research[136;137], whilst methods based on **structural equation modelling** (SEM)[80] are used in the social sciences, which include **latent growth curve modelling** (LGCM)[138;139] and **growth mixture modelling** (GMM)[140]; methods that are becoming popular in biomedical research[141]. Under certain conditions, MLM can be specified in an SEM framework using LGCM. Their similarities and differences are not just of technical interest but of practical value, revealing how flexibly one can model longitudinal data. For those familiar with MLM or LGCM, but not both, comparison of the two methods aids comprehension of the lesser known method.

For illustration, we use data from a study on the associations between a child's body growth and their mother's blood aflatoxin levels during breastfeeding in a group of 200 African children[142]. We initially examine longitudinal measures of the children's body weights and later discuss how the SEM framework allows us to relate changes in body weight (longitudinal exposure) to changes in mothers' blood aflatoxin levels (longitudinal outcome). The study data has three measures of children's body weight and their mothers' blood aflatoxin levels at birth, 3 months, and 8 months. Repeated measurements form the lowest level (level-1) of a multilevel hierarchy, with nesting at the highest level (level-2) by children for weight and by mothers for blood aflatoxin levels.

Multilevel modelling

Known also as **mixed effects modelling**, **random effects modelling**, or **hierarchical linear modelling**[143;144], this approach has mainly been used in epidemiological research to deal with hierarchical data structure (e.g. patients nested within doctors or within geographical areas), but another application of MLM is to analyse longitudinal data, treating the repeated measurements as the lowest level[137]. The basic MLM of a growth trajectory for weight, for instance, is given by:

$$Wt_{ij} = \beta_{0ij} + \beta_{1j} Time, \qquad \text{Eq.1}$$

where Wt_{ij} is body weight measured on occasion i ($i = 1, 2 \ldots T$) across T measurement time points (or T ages, e.g. 0, 3 and 8 months in the African study example), for individual j ($j = 1, 2, \ldots N$) with N total individuals (e.g. $N = 200$ in the African study dataset), and β_{0ij} / β_{1j} are multilevel regression coefficients given by:

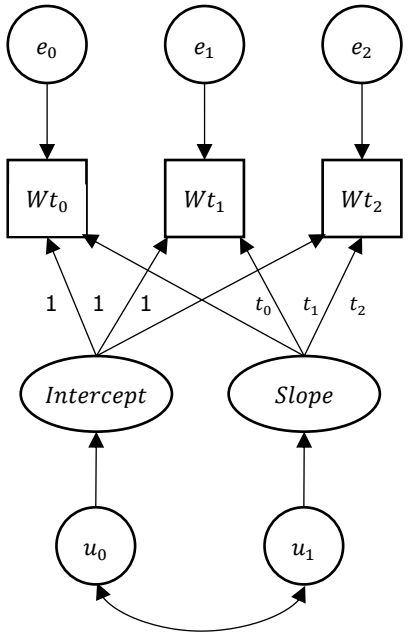
$$\beta_{0ij} = \beta_0 + u_{0j} + e_{0ij};\ \beta_{1j} = \beta_1 + u_{1j} \qquad \text{Eq.2}$$

where β_0 is the overall mean intercept (at $Time = 0$); β_1 is the overall mean gradient of the weight growth trajectory; e_{0ij} is the residual error term at level-1 representing the difference between observed and predicted weight on each occasion for each individual; u_{0j} and u_{1j} are residuals at level-2 representing, respectively, intercept and slope differences between observed mean weight trajectories for each individual and the overall mean weight trajectory for everyone. Parameters describe a population mean trajectory and how individuals deviate from that trajectory. Eq.1 assumes that growth in weight is linear and individual linear growth trajectories are estimated for everyone separately. Combined, Eq.1 and Eq.2 define a multilevel model referred to as a **random coefficient model**, since the regression coefficients exhibit random variation about their mean (across occasions and across individuals). More detailed explanations can be found elsewhere[109].

Latent growth curve modelling (LGCM)

LGCM is an application of SEM to longitudinal data[138]. Repeated measures of a variable (e.g. weight) are modelled as a function of latent factors analogous to random effects of multilevel models, with time-specific latent errors. Figure 1 is the LGCM path diagram of the MLM in Eq.1, with three observed weight variables and five latent variables: e_0 to e_2 are residual error terms for the successive measurements of body weight, u_0 and u_1 are residual errors for the two latent variables $Intercept$ and $Slope$, where u_0 and u_1 may be correlated. Each latent factor and error term is assumed to be independent and identically normally distributed.

Figure 1: SEM representation of LGCM



Numbers associated with arrows are 'factor loadings' depicting regression-like association. Loadings for residual errors are fixed to be 1, so errors are on the scale as the observed or estimated latent measures. $Intercept$ to weights loadings are unity to indicate that associations between them are equally scaled. Loadings for $Slope$ to t_0, t_1 and t_2 represent the times at which the weights were recorded.

Equivalence and differences between MLM and LGCM

For MLM and LGCM to be equivalent, two criteria must be satisfied[145]:

- The longitudinal measures must be observed at identical times for each measurement occasion; known as **interval homogeneity** (e.g. birthweights are recorded on the day of birth, weight at age 3 months is recorded say exactly 90 days after birth, and so on).
- Random slope factor loadings must reflect exact intervals between longitudinal measures (e.g. for our example factors loadings could be age in months: 0, 3, and 8; or age in months centred around the mid observational time: -4, -1 and 4; or rescaled in any way: 0.000, 0.375, 1.000).

In practice, for most longitudinal data, measurement intervals vary across individuals, even where efforts are made to minimise this (e.g. when seeking to measure weights at age 3 months, some children will be older and some younger than 90 days, with discrepancies of days or weeks). Thus, most longitudinal health data will experience **interval heterogeneity**. MLM can accommodate this easily, as time of each individual measure is the value of $Time$ in Eq.1; this may vary for everyone. In contrast, factor loadings in Figure 1 are set to be identical for everyone, and study data are then assumed to be exactly or approximately interval homogeneous for the LGCM in Figure 1.

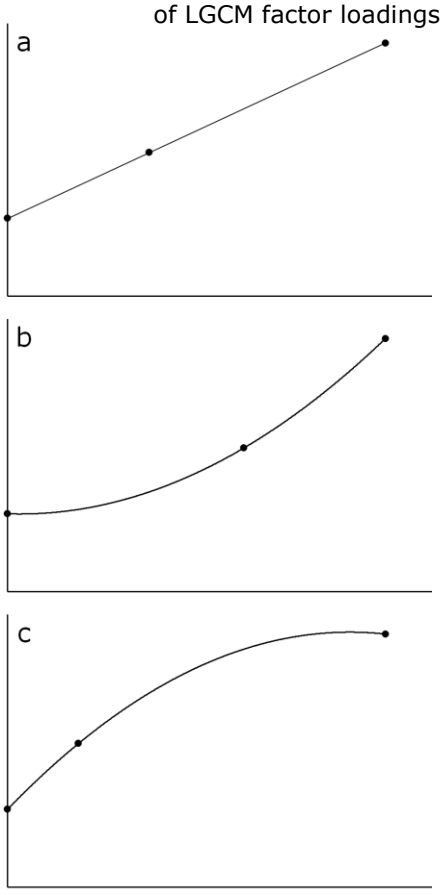
If factor loadings t_0, t_1 and t_2 are not all set, and only the first and last are set to the start and end times of the measurement period, factor t_1 is estimated as part of the modelling process.

This facilitates the modelling of nonlinear growth (even though the latent slope is linear); in effect time is 'distorted' to reflect nonlinearity.

For instance, if factors t_0 and t_2 are set to 0 and 8 (age at birth and 8 months, respectively), and if t_1 were estimated to be 3, the model would reflect linear growth (Figure 2a); but if t_1 were estimated as 5, the model then reflects nonlinear growth that is accelerating (Figure 2b); and if t_1 were estimated as 1.5, the model would reflect nonlinear growth that is decelerating (Figure 2c).

The distortion of the time axis achieved by the freely estimated factor loading for t_1 allows for nonlinear change despite only modelling a latent linear term. This flexibility of LGCM is superior to MLM for modelling nonlinearity, but relies upon all measures being interval homogeneous, which is rarely true for most longitudinal health data.

Figure 2: Graphical interpretation of LGCM factor loadings

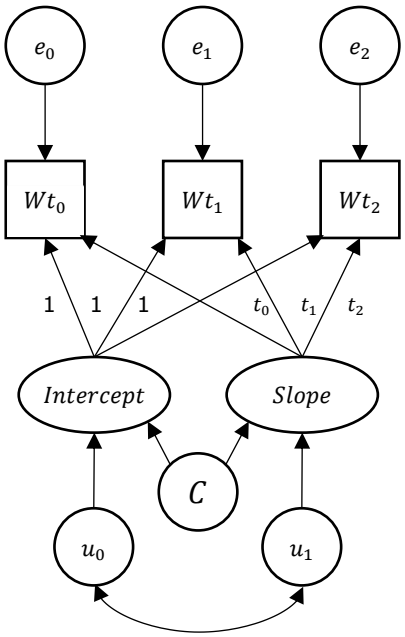


Growth mixture modelling (GMM)

GMM is an extension of LGCM where growth factors may vary across a specified number of latent classes. GMM allows for the evaluation of subgroups, and their unique patterns of change, in relation to a later-life outcome, without invoking the adverse impacts of RTM that would arise from *a priori* conditioning on the outcome or other exogenous variables. GMM is growing in popularity in biomedical research due to their potential to identify clinically meaningful subgroups, each with a specific longitudinal 'pattern' of the exposure.

Figure 3 is an extension to the SEM in Figure 1, with the latent variable C affecting the latent variables $Intercept$ and $Slope$; when $C = 1$, the model in Figure 3 is equivalent to that in Figure 1.

Figure 3: SEM representation of GMM



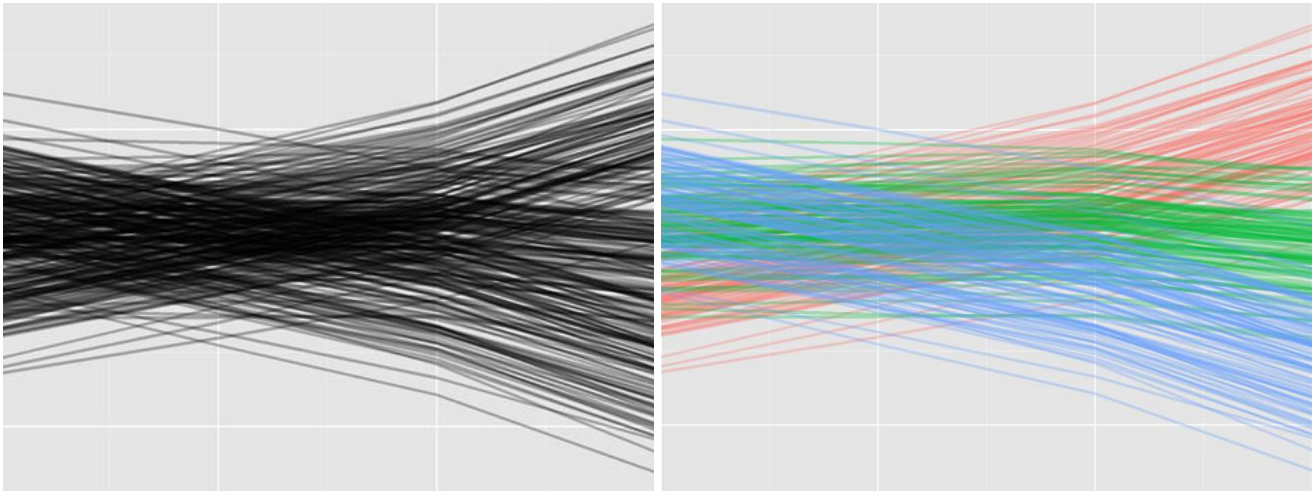
For models where $C \geq 2$, the GMM allows for identification of subgroups, each with a unique pattern of change described by separate latent $Intercept$ and $Slope$ variables for each mixture (Figure 4).

Mixtures are an inherent part of the random structure (as with ZiP/B models). In specifying C, one affects the $Intercept$ and $Slope$ estimated for each mixture, thereby influencing overall random structure in the model.

Conversely, in specifying how random effects are modelled via the latent $Intercept$ and $Slope$ variables (e.g. by constraining the correlation between u_0 and u_1 to be zero), one also influences the 'ideal' number of mixtures and their composition.

Individuals are classified by estimating posterior probabilities of class membership. If models with 2 or more classes provide a better explanation of the data than a single class model, this suggests that the population comprises subgroups with their own underlying change process. Subgroup membership is interpreted as an important feature related to later-life outcomes. Selecting a model with the 'correct' number of classes becomes central to GMM interpretation.

Figure 4: Illustration of the application of a growth mixture model to longitudinal data: (a) the individual parameterised growth curves; (b) 3 derived mixtures (modal assignment).



(a) (b)

Challenges with GMM development

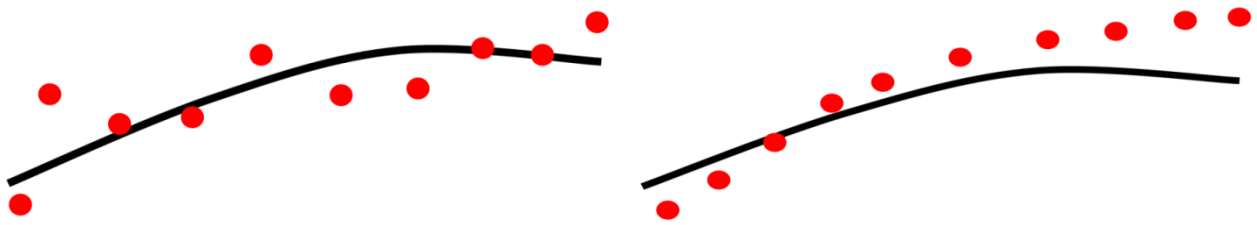
In seeking a suitable GMM, it is common practice to estimate multiple models specifying a different number of latent classes and then to decide on which model is 'best'. One approach is to constrain the factor variances of all latent classes to be zero, referred to variously as **latent class growth analysis**[146], **group-based trajectory modelling**[147], or **semi-parametric growth modeling**[148]. At the other extreme of model parsimony, one freely estimates all variance and covariance terms separately for each latent class. It is also common to select either homo- or heteroscedastic models by, respectively, constraining or freely estimating the latent error variances across time points,

with additional flexibility that the error variances are identical or different across the classes. For an increasing number of mixtures, convergence issues arise when there are too many freely estimated parameters. A common solution is to simplify the model through parameter constraints.

Arbitrary constraints may not reflect the underlying data generation process and problems arise if modifications to the random effects inadvertently introduce unwanted constraints[141]. Constraints have unintended consequences, leading to an autoregressive structure that affects the formation of mixtures and ultimately affects model interpretation based on the derived mixtures[149]. This is exacerbated if longitudinal changes *within* individuals are gradual compared to differences *between* individuals, as for most growth measures; this is not widely appreciated and since it is common to apply parameter constraints to aid convergence or to improve model parsimony, this problem is ubiquitous.

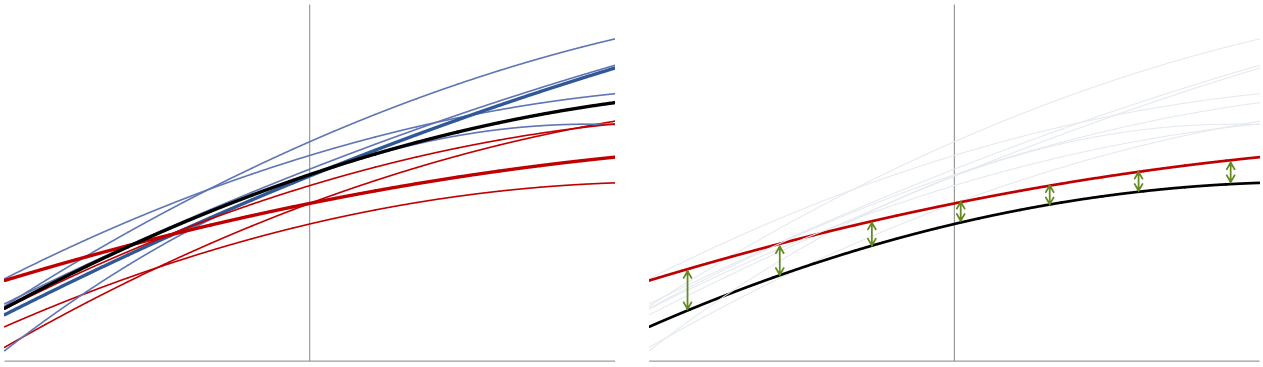
If too many variance and covariance terms are set to zero, autocorrelation emerges amongst the time-specific latent errors, since individual growth curves are consistently above or below the class-specific mean curve, and this is more likely if the exposure exhibits greater between- than within-subject heterogeneity (Figure 5). Growth measures are prone to 'tracking', i.e. where individuals that initially lie high (or low) in their centile score relative to the population distribution tend to remain high (or low) in their centile score thereafter. Although an individual's growth trajectory may cross population centiles over the longer term, for a short period at least trajectories may be relatively stable (Figure 6).

Figure 5: An illustration of individual growth (red dots) and class mean parameterised curve (black line) for: (a) greater within than between heterogeneity; (b) greater between than within heterogeneity.



(a) (b)

Figure 6: Example GMM with 2 mixtures: (a) blue and red lines depict classes (modal assignment), thick lines depict class means, black line is population mean; (b) individual (back line) 'tracks' red class mean.



(a) (b)

If variance and covariance terms must be constrained to aid model convergence, or if parsimony is preferred to aid interpretation, one strategy is to model explicitly the emergent autocorrelation structure of the random effects within mixtures[149].

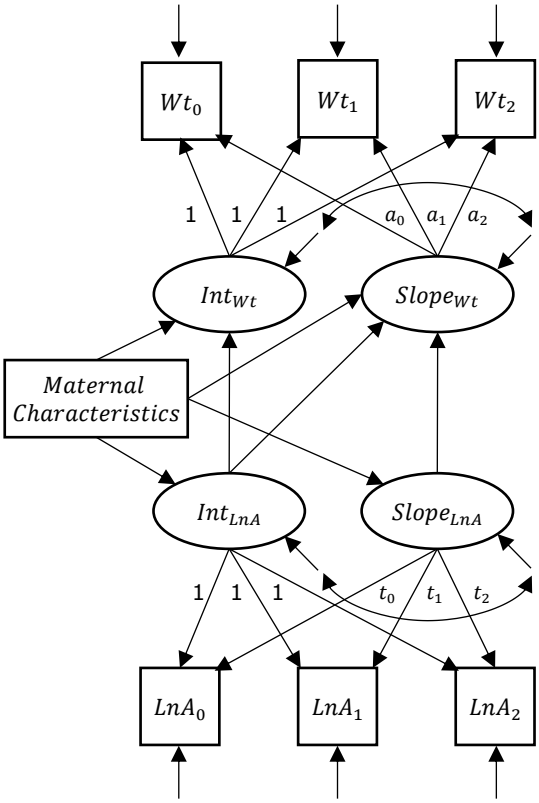
Modelling change on change

We conclude by examining how LGCM can explore the relationship between both a longitudinal exposure and a longitudinal outcome. We revisit the African study that examined growth in relation to mother's (log-transformed) blood aflatoxin levels for 200 children during breastfeeding[142] (Figure 7). Aflatoxin is a fungi-generated toxin that contaminates food world-wide[150]. In developing countries, if storage of food is not well developed, toxins contaminate grain stock and is ingested by mothers who pass them on to their children via breastfeeding[151]. It is therefore speculated that aflatoxin exposure might impair human growth and development[152;153].

We investigate this using the SEM in Figure 7, using an LGCM for both longitudinal exposure (aflatoxin blood level, log-transformed) and longitudinal outcome (child body weight). Each latent factor and error term has arrows depicting errors, though these are not drawn as latent variables. The joint model explores how initial and changing levels of mothers' aflatoxin levels affect child growth, whilst adjusting for maternal confounders (e.g. mother's BMI, age, parity, etc.).

Factor loadings a_0, a_1 and a_2 reflect children's age when weight is recorded (0, 3 and 8 months) and loadings t_0, t_1 and t_2 reflect the times when aflatoxin levels are measured (0, 3 and 8 months). We set loadings for a_0, a_1 and a_2 to 0, 0.375 and 1 (i.e. rescaled) so the latent variables Int_{Wt} and $Slope_{Wt}$ model linear weight change. As log-transformed aflatoxin levels did not change linearly[154], we set factor loadings for t_0 and t_2 to 0 and 1, leaving t_1 to be estimated, so the latent variables Int_{LnA} and $Slope_{LnA}$ model *nonlinear* change in aflatoxin exposure. Homoscedasticity was assumed.

Figure 7: Children's bodyweight modelled in relation to mother's blood levels of aflatoxin; measures at birth, 3 months and 8 months.



Summary

In the analysis of longitudinal exposures, identification of patterns or critical periods that might explain later-life outcomes is a rich and exciting area of research, yet fraught with challenges[63]. Modelling random structure is important, but needs to be considered carefully. Choice of modelling strategy, and identification of meaningful subgroups if GMM is adopted is not straightforward, as misspecification of random effects can lead to different and therefore likely incorrect conclusions. Nevertheless, the potential utility for many settings, especially with increasing availability of large and complex 'big data', makes it important to consider carefully robust strategies of modelling longitudinal observational data with methods that ensure robust causal inference.

15. STRATIFICATION ON MEDIATOR VARIABLES

Learning objectives

- Appreciate some model complexities that are feasible with latent variable models
- Learn how to evaluate mediator interactions without modelling mediators directly

Latent variable models

Multilevel models are latent variable models with a *continuous* latent variable for each upper level of the data hierarchy, for which distributional assumptions must be made (e.g. Normal). A *discrete* latent variable incorporated in a single-level model yields a mixture model, as with ZiP/ZiB models. It is possible to combine multilevel and mixture models by considering a discrete latent variable at more than one level. This permits several complex model configurations, each relating to different assumptions, with different interpretations, not all of which have analogues to continuous latent variable models or standard multilevel models; some parameterisations may not be identifiable or identifiable models may not always be interpretable.

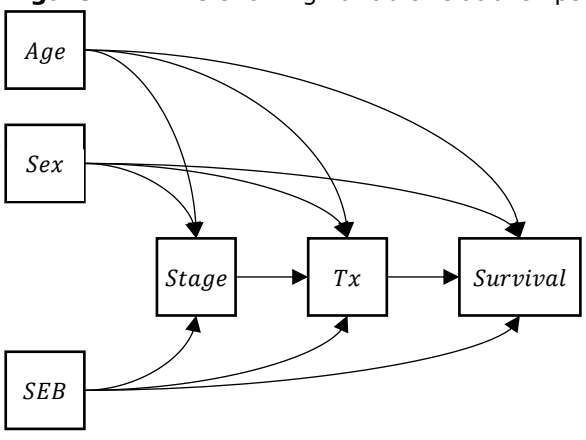
We ask: *What is the relation between 3-year (median) mortality and socioeconomic background (SEB) of patients and how does this vary with respect to tumour stage of disease at diagnosis?*

We use routinely collected data of patients registered with colorectal cancer where patients are nested within hospital Trusts. Patients with colorectal cancer (ICD-10 codes C18, C19 and C20[155]) diagnosed 1998-2004 and resident in the Northern and Yorkshire regions were identified from the Northern and Yorkshire Cancer Registry and Information Service (NYCRIS) database. Patient age, sex, tumour stage at diagnosis (using the Dukes classification[156]), diagnostic centre (Trust), and whether or not the patient received treatment were extracted. Socioeconomic background (SEB) was defined at the 2001 enumeration district level of residence (super output area) using the Townsend Index of multiple deprivation[157] and matched to patients using their postcode of residence.

The outcome is mortality (alive/dead) 3 years after diagnosis (corresponding to median survival). Patients may be treated at different Trusts throughout their care: 90% were treated in the same Trust as they were diagnosed and 75% remained with this Trust throughout. We chose to analyse the data by Trust of diagnosis to include all patients, whether treated or not, and this maintained a reasonable proportion of patients whose treatment was initially received within the same Trust as they were diagnosed. Data for 24,455 patients were available for analysis.

We seek the causal impact of SEB on 3-year survival (see DAG in Figure 1) and would like to stratify on tumour stage of disease.

Figure 1: DAG showing variable relationships



In estimating the outcome-exposure relationship in a multivariable regression model we should adjust for competing exposures *age* and *sex* (to improve model precision); however, *stage* of disease is a mediator (as too is *treatment*), which prohibits inclusion of this covariate in the regression model, thus preventing

stratification as was initially sought. To address this, we consider an alternative modelling strategy.

If stage is adjusted for in the multilevel model seeking to determine the SEB-mortality relationship, it introduces bias due to the reversal paradox[23]. With some analyses adjusting for stage and others not making this error perhaps explains why findings into the impact of SEB on cancer mortality vary; some studies find a significant relationship between worsening SEB and increased cancer mortality[158;159], whilst others find no such association[160;161]. Furthermore, regression analysis gives rise to biased results when model covariates (such as stage at diagnosis) are measured with error or have missing values[162], exacerbated within product interaction terms[163], e.g. when investigating the role of SEB across different levels of stage at diagnosis. Stage often suffers a large proportion of incomplete data. Variable quality of pathology can lead to patients being classified incorrectly[164]. There is also potential bias in the grading of stage, as the quality of pathology sometimes leads to patients being 'under-staged'[165]: for the tumour to be classified at stage C, lymph nodes must be involved, yet the number of lymph nodes retrieved is highly variable and if few nodes are available this limits the likelihood of identifying node involvement, so the tumour may instead be classified at stage B. As this impacts the treatment received, since patients diagnosed with a stage B tumour may not receive beneficial chemotherapy[166], hence the motivation to stratify any SEB-survival relationship by stage. The recording of stage has also changed over time and if a tumour is initially graded at stage C, but clinical evidence of metastatic disease is found, current policy is to 'up-stage' the tumour to stage D. Including stage as a covariate and exploring its statistical interaction with SEB thus has the potential to introduce large bias, even were the reversal paradox not of concern.

Latent variable stratification on a mediator

We explore a **multilevel latent class model** (MLLCM) that allows for subgroups of patients such that the relation between survival and SEB might vary across classes. The latent class model may include stage of disease to help differentiate classes as per any differences in stage classification. The resulting latent classes correspond to patient features that can be labelled post-hoc as per any covariate such as stage (e.g. early- or late-stage disease at diagnosis) or the outcome (e.g. 'good' or 'poor' survivors), with attention in this instance favouring the former.

When stage is included as a class predictor, and is omitted from the standard regression model, rather than as a fixed-effect covariate, resultant patient classes will yield a graduated mortality risk analogous to that observed for different stages of disease. This allows the relationship between mortality and the exposure SEB to vary across patient classes, introducing an implicit 'interaction' between stage at diagnosis and SEB, without risk of bias due to reversal paradox or measurement error on the stage covariate.

Patient classes will be derived without stage (or any other covariate) as a class predictor and a graduated differentiation across patient classes will therefore be analogous to stratification by stage of disease. In effect, this renders *stage* as a redundant covariate altogether (though not that palatable amongst those who strive hard to improve the coding quality of staging!).

Discussion

When investigating the relationship between patients' socioeconomic circumstances and cancer mortality, individual measures of deprivation are rarely available, especially when using routine

data. Indices of SEB, such as the Townsend Index[157] and the Index of Multiple Deprivation[167], are all that is usually available. These indices are measured at the small-area level, such as electoral ward or super-output area. This can lead to the ecological fallacy[168] if area-based findings are extrapolated to individuals living in each area. For this reason, another level should be introduced (the small-area level) and this would be 'cross-classified' with Trusts, i.e. patients from one small area might attend different Trusts and similarly patients from one Trust may be drawn from different small areas of residence. Similarly, instead of the binary outcome, survival analysis (e.g. using Cox proportional hazards regression) would be used instead. All these more complex model extensions are possible.

By not modelling stage as a mediator, we can avoid the reversal paradox and minimise bias due to measurement error and/or incomplete data, but stratification on this variable is then not available using standard regression methods. With stage included as a class predictor, bias due to the reversal paradox is certainly reduced though may not be completely eradicated. However, as patient classes may be derived without stage as a class predictor, if similar differentiation across patient classes is observed, then stage may be deemed redundant; if this is not palatable to those who prefer to use this variable, then at least the latent class approach will have reduced bias than the traditional approach.

16. A CAUTIONARY NOTE ON STATISTICAL INTERACTION

Learning objectives

- Appreciate the distinction between *statistical interaction* and *biological joint effects*
- Be aware of the importance of linearity and scale in multivariable regression
- Appreciate the importance of effect size over significance testing for statistical interaction
- Be aware of the futility of most power calculations for statistical interaction

Statistical vs. biological interaction

Statistical interactions are often used in multivariable regression models to explore the joint effects of putative causal agents in relation to a single outcome, yet the statistical process is frequently misunderstood or misinterpreted. Different language used, such as effect modification or even confounding, is misleading. Effect modification is an explicit parametrisation of causal action that need not be entirely linked to a single statistical interaction, whilst confounding is a distinct concept that may or may not involve statistical interaction[169]. Consequently, attempts are made to interpret statistical interaction as though representing biological interaction, yet the two concepts need not be linked[170-173]. Statistical interaction is a well-defined mathematical concept, yet its interpretational issues lie in the contrast between parametric realisations bestowed by a statistical model upon the implied underlying stochastic nonparametric causal mechanisms.

Distinction between model parameterisation and causal process is critical. Our world unfolds based on immutable physical and biological laws that cannot be transformed, merely observed and described through experimentation. Statistical models may represent these experiments, though model construct is a matter of choice, with parameterisations often adopted out of convenience. It is important to reflect upon biological effects in a causal framework and to explore contexts in which this is meaningfully summarised by a multivariable model; *explicit* interpretation of biological mechanisms follow only if the model has a direct biological analogue, else there is at best *implicit* biological interpretation. Care must be taken relating biological processes to aspects of a statistical model and vice versa. This is particularly well illustrated in the exploration of joint effects of genes and environmental risk factors on disease, though generalises to all forms of statistical interaction.

Mapping biological process onto a statistical model

Genetic and environmental effects may operate mechanistically in different ways. For instance, a genetic polymorphism may 'program' a condition to occur absolutely, e.g. cystic fibrosis occurs definitively as a consequence of the CFTR polymorphism on chromosome seven[174]. Alternatively, individuals might merely have a greater predisposition of developing a condition, e.g. deep vein thrombosis (DVT) is more likely but not definitively a consequence of Factor V Leiden genetic mutation on chromosome one[175], whilst DVTs also occur amongst normal individuals. Mechanisms by which an outcome occurs likely involves multiple stages in biology, of which some are necessary and sufficient, whilst others modify the likelihood of occurrence. It is thus necessary to distinguish between explicit causal mechanisms that are understood biologically and implicit causality that is an abstraction or oversimplification of the more complex real world. Both may be described by a DAG and evaluated statistically within an appropriate model, but the latter DAG need not map onto any biological process precisely, which is important for the causal interpretation of joint action.

Consider DVT and exposure to the combined oral contraceptive pill (COCP). Genetically normal individuals develop DVT, though risk is elevated amongst individuals with Factor V Leiden genetic mutation[175]. Amongst women, exposure to the COCP yields an elevated risk of DVT[176]. Considering the joint action of genetic mutation and combined pill, the putative causal process is best captured by an underlying *risk* of developing DVT, i.e. a continuous probability between zero and one: the genetic and environmental exposures operate jointly to affect the *risk* of DVT, yet both exposures and the outcome are typically taken to be binary. This has implications on model interpretation. A statistical model depicts this via a logistic model. Simplification of the underlying biology is fine, but what becomes of the *biological* interpretation of the gene-environment statistical interaction?

Model parameterisation

There is no concept of statistical interaction (or even linearity) within the nonparametric causal framework of a DAG; these concepts enter centre stage only when we build our model. The choice available to us when employing statistical regression is often driven by mathematical convenience, not underlying biological processes. Statistical interaction is both *linear* and *scale-dependent*; these features are critical in meaningful interpretation of statistical interaction.

A **linear model** is defined as linear *in the fitted coefficients*, which implies that the coefficients of the model are additive. A **nonlinear model** has coefficients that *are not combined additively*. We focus only on linear models. Confusion can occur when a functional relationship, known as the *link* function, operates on an outcome. Such functions ensure that the right-hand side remains linear and these models are known as generalised linear models (GLMs). For instance, logistic regression, a key analytical tool in epidemiology, uses the *logit* link. Most statistical methods evaluating genetic and environmental factors affecting disease outcomes are likely to be a linear logistic regression model using the *logit* link function.

Statistical interaction in a linear regression model has the form of a *product term*, describing deviation from the additive effects of the product components on some predefined link function scale. In the simple case of a linear regression model with a continuous outcome (i.e. identity link) and two covariates, say *treatment group* and *sex* (both binary), the interaction term is the product of *treatment group* with *sex*. This product term is included in the model to allow treatment effects to differ for males and females. In this instance, some say that sex 'modifies' the effect of the treatment (hence the term *effect modification*), though we should be cautious about language that implies cause and effect (unless intended). It is not possible to display the parametric concept of statistical interaction in a DAG. Since an interaction allows for comparison of treatment effects for males and females, the term *subgroup analysis* is also used. The interpretation of main effects is different in the presence of an interaction term: each main effect refers to the reference group of other variables and to obtain the effect in the non-reference group, all three estimates (*treatment*, *sex* and *treatment*.*sex*) must be considered simultaneously in combination.

The importance of scale in statistical interaction

The choice of link function (e.g. *identity* vs. *logit*) affects the scale upon which covariate changes are associated with the outcome. Switching between a continuous model (*identity* link) and a binary model (*logit* link) changes model scale from *additive* to *multiplicative*. Consider the outcome *Blood Pressure* (*BP*) measured in millimetres' mercury (mmHg), dichotomised across the threshold of

140 mmHg to create a binary outcome *Hypertension* (*Hyp*). Also consider two covariates: a genetic binary variable (*G*) to depict individuals with a genetic mutation predisposing to hypertension (coded 1 if present, 0 otherwise); and an environmental variable (*E*) recorded as a binary to depict high or low salt intake (coded 1 or 0 respectively); and assume that both the genetic mutation and high salt intake elevates blood pressure.

The *normal* linear model is:

$$BP = \beta_0 + \beta_1 G + \beta_2 E + \beta_3 GE + e \qquad \text{Eq.1}$$

The *binary* logistic model is:

$$\text{logit}(Hy) = \ln\left(\frac{p}{1-p}\right) = \gamma_0 + \gamma_1 G + \gamma_2 E + \gamma_3 GE + \epsilon \ . \qquad \text{Eq.2}$$

For no statistical interaction, $\beta_3 = 0$ in Eq.1 and $\gamma_3 = 0$ in Eq.2; for a synergistic interaction, $\beta_3 > 0$ and $\gamma_3 > 0$; for an antagonistic interaction, $\beta_3 < 0$ and $\gamma_3 < 0$. Chart model coefficients two ways: (a) using a single chart for all model coefficients showing their relative effect sizes regarding the genetic wild type and low salt intake group; or (b) & (c) separate charts for high and low salt intake, respectively, contrasting genetic mutation to the wild type (i.e. the 'normal' genetic form).

With no statistical interaction between genetic mutation and salt intake, the difference in blood pressure within the *normal* model between those with and without the genetic mutation is 10 mmHg and the difference between those with and without high salt intake is 15 mmHg, seen in both chart formats (Figure 1). With an antagonistic statistical interaction, there is a smaller elevated blood pressure (20 mmHg) for the combined genetic mutation and higher salt intake than expected from adding the separate effects of genetic mutation and high salt intake (10 mmHg + 15 mmHg ≠ 20 mmHg), and this is observed in both graphical formats (Figure 2).

Considering the *logistic* model with no statistical interaction, plotting odds ratios (ORs: exponential of the model coefficients), the absolute difference in the wild type versus mutation odds ratios for hypertension between low and high salt intake is $(14.0 - 4.9) - (2.8 - 1.0) = 7.3$ (Figure 3a), not zero. When the ORs for elevated blood pressure are plotted separately for low and high salt intake, their absolute difference *is* zero (Figures 3b & 3c: $(2.8 - 1.0) - (2.8 - 1.0) = 0.0$). For an antagonistic statistical interaction, the combined chart (Figure 4a) reveals a small absolute difference $[(10.5 - 6.4) - (3.9 - 1.0) = 1.2]$ in the wild type versus mutation ORs for hypertension between low and high salt intake, whilst the separate charts for low and high salt intake indicate a larger absolute difference $[(3.9 - 1.0) - (1.6 - 1.0) = 2.3]$, nearly twice as large (Figures 4b & 4c).

How model coefficients are plotted, i.e. in combined or separate charts, gives rise to different 'visual' indications of the presence / absence of a statistical interaction and its magnitude; only the two-chart format is correct (though this is perhaps impractical for regular use). It is the graphical scale adopted that creates confusion, as the normal model should be (and is) displayed on the additive scale, whilst the logistic model should be (but was not) displayed on the multiplicative odds ratio scale; the combined chart is informative of statistical interaction only on the correct scale. Thus, the log scale is adopted for displaying odds ratios. In general, a chart's y-axis must be transformed by the link function to avoid misleading graphical display of model coefficients regarding the presence, absence or magnitude of a statistical interaction.

Figure 1: A normal model for hypertension without statistical interaction between genotype and salt intake, showing coefficients combined and separately for low and high salt intake

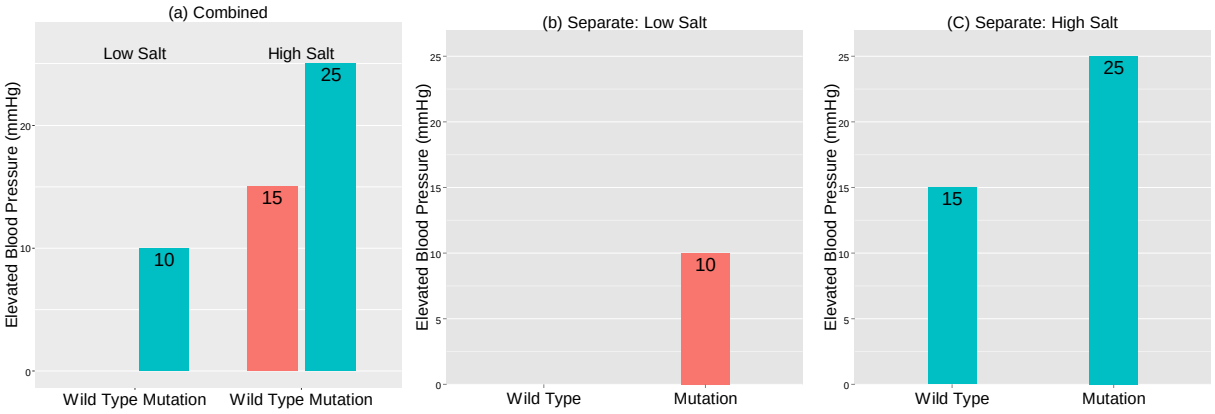


Figure 2: A normal model for hypertension with antagonistic statistical interaction between genotype and salt intake, showing coefficients combined and separately for low and high salt intake

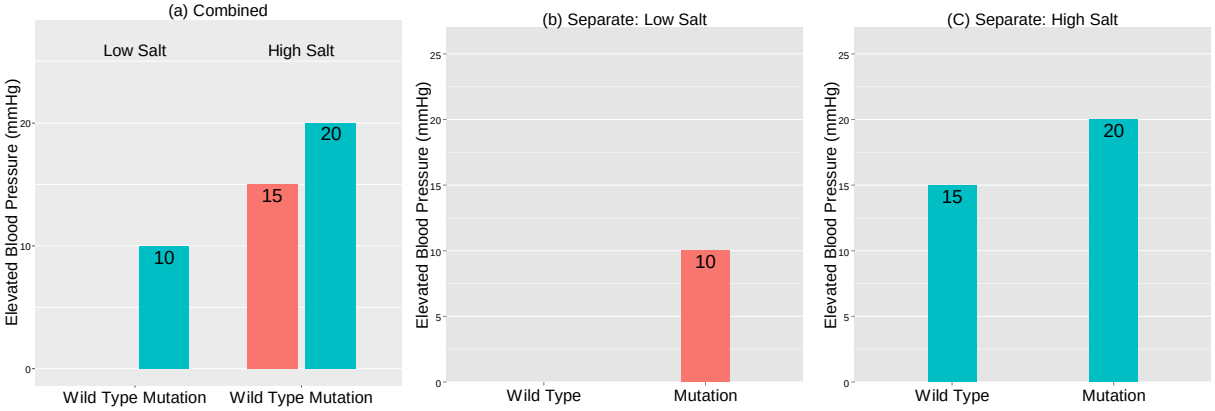


Figure 3: A logistic model for hypertension without statistical interaction between genotype and salt intake, showing coefficients combined and separately for low and high salt intake

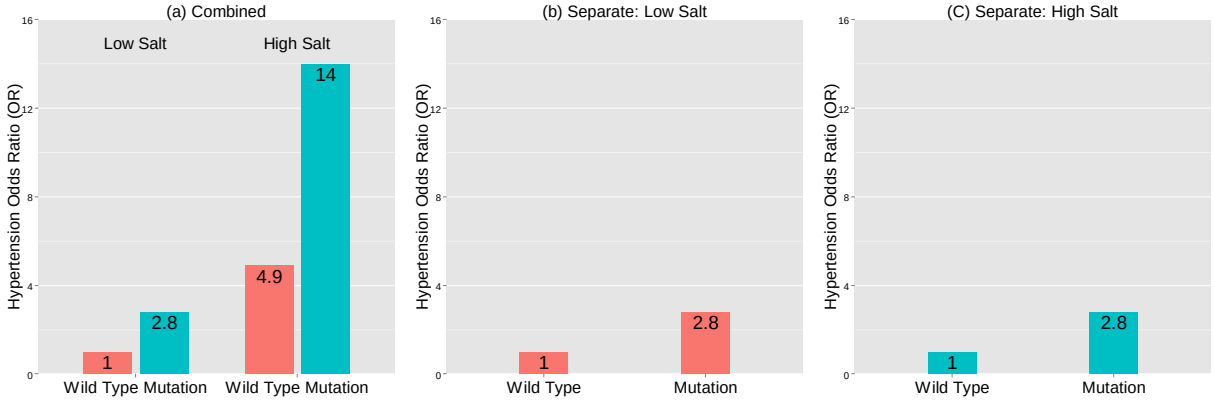
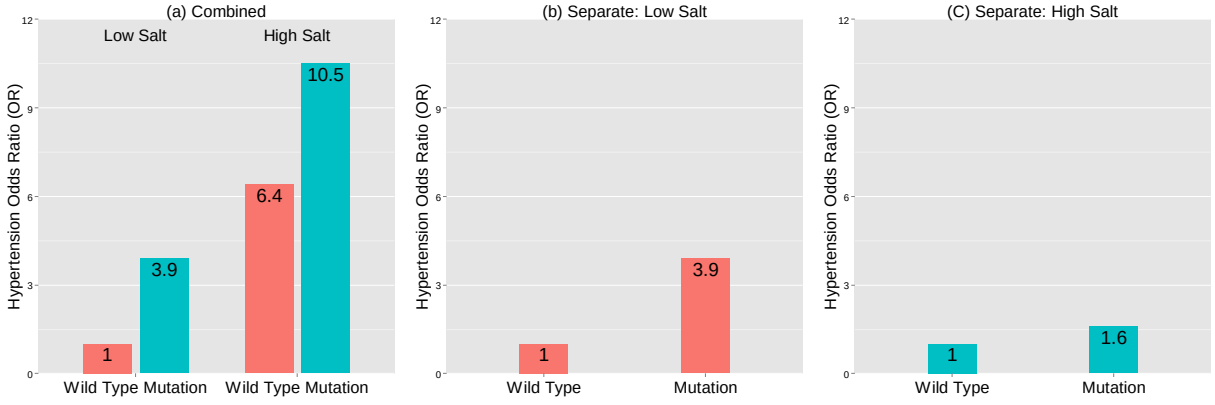


Figure 4: A logistic model for hypertension with antagonistic statistical interaction between genotype and salt intake, showing coefficients combined and separately for low and high salt intake



All regression models are scale dependent, which matters when seeking to interpret statistical interaction. There is nothing special about the scales adopted by most models; they are usually chosen for statistical convenience. There are an infinite number of possible scales on which model covariates could relate to the outcome, depending on the link function chosen. As there are only a handful of regularly used link functions, it is easy to overlook how arbitrary model scale is.

The importance of the linearity in statistical interaction

Consider the following linear model:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 z + \beta_4 xz + e \qquad \text{Eq.3}$$

where y is a continuous outcome, β_0 is the intercept, x is the exposure of interest and known to exhibit a curvilinear relationship with the outcome (hence the quadratic term in x), z is a continuous confounder, β_i (i=1…4) are covariate regression coefficients, and e is residual error that is normally distributed with mean zero and variance σ^2.

In the parametric development of our causal thinking (i.e. transitioning from DAG to multivariable model), we anticipate an interaction between x and z, hence the product interaction term xz. Were we to find there is no xz interaction, then $\beta_4 = 0$ and the correct model would be:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 z + e. \qquad \text{Eq.4}$$

Now if we overlooked the curvilinear relationship, i.e. dropped the quadratic term (x^2) in Eq.3, whilst still exploring the possible xz interaction, the model we would consider is:

$$y = \beta_0 + \beta_1 x + \beta_3 z + \beta_4 xz + e \qquad \text{Eq.5}$$

and a 'spurious' interaction (i.e. $\beta_4 \neq 0$) is likely to be observed in Eq.10 for the xz product interaction[177]. The collinearity between x and z effectively 'mops up' the unaccounted outcome variance that would have been accommodated by the curvilinear relationship between y and x, and the statistical interaction is observed. The assumption of linearity between the outcome y and covariate x that is not upheld gives rise to the apparent statistical interaction. Over-simplification of statistical models in contrast to the complex biological processes they emulate, plus the arbitrary choice of link function, makes it unsurprising that several covariate-outcome relationships are nonlinear, with implications for statistical interactions.

If underlying nonlinear relationships are overlooked in a multivariable model, statistical interaction will be observed. This is not 'spurious', as the model is mathematically sound; the issue is one of interpretation. The basis of a statistical interaction may be entirely statistical, not biological, and any causal interpretation of the statistical interaction may be misguided. Complex parameterisation (i.e. nonlinearity and/or interaction) rarely bestows insight regarding the putatively causal (joint) action biologically; parameterisation is an extension within the multivariable modelling toolkit from that of the simplest default starting point of a linear model where all relationships are assumed to be linear and have no interactions.

Statistical power to test for interaction

It is well documented that much larger sample sizes are required to test statistical interactions than for main effects[178], and this is a criticism directed at many studies often berated for being too small to examine gene-environment interactions. Notwithstanding the overzealous nature to test

(and the associated undesirable addiction to *p*-values), the reasoning behind such sample-size criticisms is also flawed because the statistical power of an interaction is just as scale-dependent as the statistical interaction itself. Without a meaningful scale upon which the test is sought, there is no basis for power calculations. Perversely, one could take data from pilot studies and transform the data by trial and error to test different models repeatedly until one finds a transformation (hence a variable scale) upon which the sample size estimated is minimal. This strategy could save enormous expenditure in epidemiology, were it not as erroneous as the focus on p-values.

Causal interpretation

Within a causal inference framework, to describe the *magnitude* of causal effects, we are compelled to appreciate the causal nature of the variables involved, as within a DAG. The model we use must also reflect plausible parametric relationships amongst our observed variables. Compressing the complexity of biological processes into the simplicity of a statistical model is often unrealistic, and we should not seek to infer detailed understanding of biology from statistical models. Instead, we should have an *a priori* overview of plausible causal mechanisms and use statistical modelling to estimate causal effect sizes that have clinical meaning.

Return to the illustration for deep vein thrombosis (DVT), Factor V Leiden genetic mutation and the environmental combined oral contraceptive pill (COCP). We might seek the joint association of genetic mutation and COCP with respect to DVT. Acknowledging that the environmental exposure (COCP) is not strictly categorical (COCP exposure varies per dose and by the extent of use), we nevertheless categorise this into present or absent, i.e. whether a woman uses the COCP or not. We might then examine any statistical interaction for the data summarised in Table 1 from a case-control study[175]. Point estimates are presented, though 95% CIs should also be calculated.

Table 1: Summary of the case-control study investigating the joint association of both Factor V Leiden genetic mutation and the combined oral contraceptive pill (COCP) use with respect to deep vein thrombosis (DVT)

Factor V / COCP	Cases	Controls	OR
+/+	25	2	34.7
+/-	10	4	6.9
-/+	84	63	3.7
-/-	36	100	1.0
Totals	155	169	

As the analysis is undertaken using odds ratios, it is appropriate to consider the *multiplicative* scale when interpreting joint effects. To examine **departure from a *multiplicative* model** (i.e. with no statistical interaction) we take the ratio of observed (34.7) and expected ($3.7\times6.9\times1.0=25.7$) odds ratios, i.e. $34.7/25.7=1.4$, and contrast this to unity (null effect on the odds ratio scale). The departure of 1.4 from 1.0 is small, but may be statistically significant for large studies, suggesting that there is a hint of statistical interaction on the OR (multiplicative) scale.

We can examine **departure from an *additive* model** (i.e. with no statistical interaction), taking the difference between observed (34.7) and expected ($3.7+6.9-1.0=9.6$) odds ratios, i.e. $34.7-9.6=25.1$, and contrasting to zero (null effect on the standard linear scale). We note that 25.1 is far from zero and likely to be statistically significant for all but very small studies, suggesting strong evidence of statistical interaction on the standard linear (additive) scale.

What insight is gained from formally *testing* either departure if we do not know how to interpret statistical interaction (if present) on either scale? It is apparent that the multiplicative model fits the data closer than the additive model, statistically speaking, suggesting that the 'joint effects' of the genetic mutation and environmental exposure are approximately multiplicative on the OR scale, but from a causal perspective, what does this tell us? What does the information in Table 1 mean with respect to public health: how does it inform women considering the COCP?

Relative to *not* having Factor V Leiden genetic mutation and *not* using the COCP, taking the contraceptive increases a woman's relative risk (RR) for DVT by approximately 3.6-fold (since DVT is rare, hence OR ≈ RR). If there was no reason for the woman to suspect she had the genetic mutation, which has a prevalence of around 4.4% in Europe[179], these increased risks may or may not worry her. On the other hand, if she was aware of a family history of DVT, she may fear that she carries the genetic mutation, and considering the relative risk of having *both* the mutation *and* using COCP (RR≈34.7) compared to merely having the mutation (RR≈6.9), she might then seek to use alternative contraception, or explore being genetically tested first. Whilst there is no knowing how a woman would choose to use the information in Table 1, it is dubious to suppose her interest lies in the *p*-value of a formal test to verify a strong *synergistic statistical interaction* on the *additive scale*, or to establish the absence of an interaction on the multiplicative OR scale. The framework in which each woman's decision is formed is likely influenced by the relative risk *effect size* (along with its confidence interval) than any formal test. Thus: *Why focus on formal tests for statistical interaction?* A related concern is: *Why focus on the statistical power of such tests?*

Summary

Invoking causal inference of joint processes in a statistical model may be at best *consistent* with some form of biological analogue; testing for statistical interaction does not contribute to causal understanding. Quantification of joint effects remains a legitimate goal, but rarely does its utility lie in the elucidation of *biological* process[180]. Overzealous interpretation of statistical interaction has the potential to invoke misunderstanding of the causal mechanisms operating. Statistical interactions are meaningful only in regards their estimated effect size and associated 95% CI and this effect size can be manipulated by either categorisation of continuous measures or application of other transformations. The obsession to test for statistical interaction is misguided and fuels attention to study sample size, pressuring researchers to seek sufficient statistical power for the elucidation of *statistically significant* joint effects. Consequently, there is a perceived and falsely legitimised demand for increasingly large studies. Insufficient attention is given to these issues[181]. Less attention is given to estimating the effect size of joint effects for *clinical interpretation*. Such practices are commonplace in the pursuit of gene-environment interactions, though the very same issues apply to all aspects of epidemiology (and beyond). The key is always to begin with a causal framework, and only then engage with parameterisation of models appropriate for that framework, and not the other way around.

REFERENCES

1. Flom PL, Cassell DL. Stopping stepwise: Why stepwise and similar selection methods are bad, and what you should use. Proceedings of the North East SAS Users Group Inc 20th Annual Conference 2007; http://www.nesug.org/proceedings/nesug07/sa/sa07.pdf.
2. Pickering TG, James GD, Boddie C, Harshfield GA, Blank S, Laragh JH. How common is white coat hypertension? JAMA 1988; 259(2):225-228.
3. Young SS, Karr A. Deming, data and observational studies: A process out of control and needing fixing. Significance 2011; 8(3):116-120.
4. Stefanski LA. The effects of measurement error on parameter estimation. Biometrika 1985; 72(3):583-592.
5. Gatignon H. Error in Variables: Analysis of Covariance Structure – Structural Equation Models. Statistical Analysis of Management Data. US: Springer; 2014. 297-348.
6. Rubin DB. Estimating causal effects from large data sets using propensity scores. Ann Intern Med 1997; 127(8 Pt 2):757-763.
7. Martens EP, Pestman WR, de BA, Belitser SV, Klungel OH. Instrumental variables: application and limitations. Epidemiology 2006; 17(3):260-267.
8. Angrist JD, Imbens GW, Rubin DB. Identification of Causal Effects Using Instrumental Variables. Journal of the American Statistical Association 1996; 91(434):444-455.
9. Greenland S, Robins JM, Pearl J. Confounding and Collapsibility in Causal Inference. Statistical Science 1999; 14:29-46.
10. Hoggart CJ, Parra EJ, Shriver MD, Bonilla C, Kittles RA, Clayton DG et al. Control of confounding of genetic associations in stratified populations. Am J Hum Genet 2003; 72(6):1492-1504.
11. Hernan MA, Hernandez-Diaz S, Robins JM. A structural approach to selection bias. Epidemiology 2004; 15(5):615-625.
12. Rothman KJ. The estimation of synergy or antagonism. Am J Epidemiol 1976; 103(5):506-511.
13. Textor J, van der Zander B, Gilthorpe MS, Liskiewicz M, Ellison GT. Robust causal inference using directed acyclic graphs: the R package 'dagitty'. Int J Epidemiol 2017.
14. Greenland S, Robins JM. Identifiability, exchangeability, and epidemiological confounding. Int J Epidemiol 1986; 15(3):413-419.
15. McNamee R. Confounding and confounders. Occup Environ Med 2003; 60(3):227-234.
16. Greenland S, Morgenstern H. Confounding in health research. Annu Rev Public Health 2001; 22:189-212.
17. Hennekens CH, Buring JE. Epidemiology in Medicine. Boston: Little, Brown & Co; 1987.
18. Pearl J. Causality: Models, Reasoning, and Inference. 2nd ed. Cambridge: Cambridge University Press; 2009.
19. Belsky DW, Caspi A, Houts R, Cohen HJ, Corcoran DL, Danese A et al. Quantification of biological aging in young adults. Proc Natl Acad Sci U S A 2015; 112(30):E4104-E4110.
20. Schaefer JD, Caspi A, Belsky DW, Harrington H, Houts R, Israel S et al. Early-Life Intelligence Predicts Midlife Biological Age. J Gerontol B Psychol Sci Soc Sci 2015.
21. Box GEP, Draper NR. Empirical Model-Building and Response Surfaces. New York: Wiley; 1987.
22. Tu YK, Gunnell D, Gilthorpe MS. Simpson's Paradox, Lord's Paradox, and Suppression Effects are the same phenomenon - the reversal paradox. Emerg Themes Epidemiol 2008; 5:2.
23. Tu YK, West R, Ellison GT, Gilthorpe MS. Why evidence for the fetal origins of adult disease might be a statistical artifact: the "reversal paradox" for the relation between birth weight and blood pressure in later life. Am J Epidemiol 2005; 161(1):27-32.
24. Pearl J. Comment: Understanding Simpson's Paradox. The American Statistician 2014; 68(1):8-13.
25. Tu YK, Ellison GT, West R, Gilthorpe MS. Tu et Al. Respond to "barker meets simpson". Am J Epidemiol 2005; 161(1):36-37.
26. Tu YK, Ellison GT, Gilthorpe MS. Growth, current size and the role of the 'reversal paradox' in the foetal origins of adult disease: an illustration using vector geometry. Epidemiol Perspect Innov 2006; 3:9.
27. Tu YK, Gilthorpe MS, Ellison GT. What is the effect of adjusting for more than one measure of current body size on the relation between birthweight and blood pressure? J Hum Hypertens 2006; 20(9):646-657.
28. Tu YK, Manda SO, Ellison GT, Gilthorpe MS. Revisiting the interaction between birth weight and current body size in the foetal origins of adult disease. Eur J Epidemiol 2007; 22(9):565-575.
29. Weinberg CR. Invited commentary: Barker meets Simpson. Am J Epidemiol 2005; 161(1):33-35.
30. Pearl J. On Simpson's Paradox. Again? http://www.mii.ucla.edu/causality/?p=1189: 2014.
31. Pearl J. Causal diagrams for empirical research. Biometrika 1995; 82(4):669-688.

32. Ellison GT, de WT. Poverty, disability and self-reported health amongst residents and migrants in Gauteng, South Africa. Ann Hum Biol 2016; 43(2):131-143.
33. Westreich D, Greenland S. The table 2 fallacy: presenting and interpreting confounder and modifier coefficients. Am J Epidemiol 2013; 177(4):292-298.
34. Pearl J. Do-Calculus Revisited. In: Freitas Nde, Murphy K, editors. Proceedings of the Twenty-Eighth Conference on Uncertaintyin Artificial Intelligence. Corvallis, OR: AUAI Press; 2012. 4-11.
35. Pearl J. Interpretation and identification of causal mediation. Psychol Methods 2014; 19(4):459-481.
36. Pearl J, Bareinboim E. External validity: From do-calculus to transportability across populations. Statistical Science 2014; 29(4):579-595.
37. Huxley RR, Neil A, Collins R. Unravelling the fetal origins hypothesis: is there really an inverse association between birthweight and subsequent blood pressure? Lancet 2002; 360(9334):659-665.
38. Lippa RA. Gender, Nature, and Nurture. New York: Taylore & Francis; 2005.
39. Halpern DF. Sex Differences in Cognitive Abilities. Fourth ed. New York: Taylor and Francis; 2012.
40. Hernandez-Diaz S, Schisterman EF, Hernan MA. The birth weight "paradox" uncovered? Am J Epidemiol 2006; 164(11):1115-1120.
41. Wilcox AJ. Invited commentary: the perils of birth weight--a lesson from directed acyclic graphs. Am J Epidemiol 2006; 164(11):1121-1123.
42. Wadsworth ME, Hardy RJ, Paul AA, Marshall SF, Cole TJ. Leg and trunk length at 43 years in relation to childhood health, diet and family circumstances; evidence from the 1946 national birth cohort. Int J Epidemiol 2002; 31(2):383-390.
43. Tu YK, Woolston A, Baxter PD, Gilthorpe MS. Assessing the impact of body size in childhood and adolescence on blood pressure: an application of partial least squares regression. Epidemiology 2010; 21(4):440-448.
44. Hu FB, Stampfer MJ, Rimm E, Ascherio A, Rosner BA, Spiegelman D et al. Dietary fat and coronary heart disease: a comparison of approaches for adjusting for total energy intake and modeling repeated dietary measurements. Am J Epidemiol 1999; 149(6):531-540.
45. Ng SW, Popkin BM. Time use and physical activity: a shift away from movement across the globe. Obes Rev 2012; 13(8):659-680.
46. Davey Smith G, Greenwood R, Gunnell D, Sweetnam P, Yarnell J, Elwood P. Leg length, insulin resistance, and coronary heart disease risk: the Caerphilly Study. J Epidemiol Community Health 2001; 55(12):867-872.
47. Yudkin PL, Stratton HH. How to deal with regression to the mean in intervention studies. Lancet 1996; 347:241-243.
48. Fitzmaurice G. A conundrum in the analysis of change. Nutrition 2001; 17(4):360-361.
49. Morton V, Torgerson DJ. Effect of regression to the mean on decision making in health care. BMJ 2003; 326(7398):1083-1084.
50. Bland JM, Altman DG. Regression towards the mean. British Medical Journal 1994; 308(6942):1499.
51. Bland JM, Altman DG. Some examples of regression towards the mean. British Medical Journal 1994; 309(6957):780.
52. Galton F. Regression towards mediocrity in hereditary stature. The Journal of the Anthropological Institute of Great Britain and Ireland 1886; 15:246-263.
53. Campbell DT, Kenny DA. A primer on regression artefact. Guildford: The Guilford Press; 1999.
54. Michels KB, Trichopoulos D, Robins JM, Rosner BA, Manson JE, Hunter DJ et al. Birthweight as a risk factor for breast cancer. Lancet 1996; 348(9041):1542-1546.
55. Stavola BL, Hardy R, Kuh D, Silva IS, Wadsworth M, Swerdlow AJ. Birthweight, childhood growth and risk of breast cancer in a British cohort. Br J Cancer 2000; 83(7):964-968.
56. Andersen ZJ, Baker JL, Bihrmann K, Vejborg I, Sorensen TI, Lynge E. Birth weight, childhood body mass index, and height in relation to mammographic density and breast cancer: a register-based cohort study. Breast Cancer Res 2014; 16(1):R4.
57. Sayers A, Heron J, Smith ADAC, Macdonald-Wallis C, Gilthorpe MS, Steele F et al. Joint modelling compared with two stage methods for analysing longitudinal data and prospective outcomes: A simulation study of childhood growth and BP. Statistical Methods in Medical Research 2014;in press.
58. Bhargava SK, Sachdev HS, Fall CH, Osmond C, Lakshmy R, Barker DJ et al. Relation of serial changes in childhood body-mass index to impaired glucose tolerance in young adulthood. N Engl J Med 2004; 350(9):865-875.
59. Barker DJ, Osmond C, Forsen TJ, Kajantie E, Eriksson JG. Trajectories of growth among children who have coronary events as adults. N Engl J Med 2005; 353(17):1802-1809.
60. Kahn SE, Hull RL, Utzschneider KM. Mechanisms linking obesity to insulin resistance and type 2 diabetes. Nature 2006; 444(7121):840-846.

61. Barker DJ, Gluckman PD, Godfrey KM, Harding JE, Owens JA, Robinson JS. Fetal nutrition and cardiovascular disease in adult life. Lancet 1993; 341(8850):938-941.
62. Gluckman PD, Hanson MA, Buklijas T. A conceptual framework for the developmental origins of health and disease. J Dev Orig Health Dis 2010; 1(1):6-18.
63. Tu YK, Tilling K, Sterne JA, Gilthorpe MS. A critical evaluation of statistical approaches to examining the role of growth trajectories in the developmental origins of health and disease. Int J Epidemiol 2013; 42(5):1327-1339.
64. Murray ET, Hardy R, Hughes A, Wills AK, Sattar N, Deanfield JE et al. How is overweight/obesity across the life course associated with levels of adipokines, inflammatory and endothelial markers at age 60-64 years? Findings from the 1946 birth cohort. J Epidemiol Community Health 68, A34-A35. 2014.
65. Berrie L, Baxter PD, Norman PD, Ellison GTH, Law GR, Feltbower RG et al. Different analytical strategies yield contradictory findings when investigating the association between childhood leukaemia and population mixing. Journal of Epidemiology & Community Health 2016; 70(Supple 1):A87.
66. Kinlen L. Evidence for an infective cause of childhood leukaemia: comparison of a Scottish new town with nuclear reprocessing sites in Britain. Lancet 1988; 2(8624):1323-1327.
67. Pearson K, Lee A, Elderton EM. On the correlation of death-rates. Journal of the Royal Statistical Society 1910; 73:534-539.
68. Neyman J. Human cancer: radiation and chemicals compete. Science 1979; 205:259-260.
69. Fisher RA. The analysis of covariance method for the relation between a part and the whole. Biometrics 1947; 3(2):65-68.
70. Clark BR, Ferketich AK, Fisher JL, Ruymann FB, Harris RE, Wilkins JR, III. Evidence of population mixing based on the geographical distribution of childhood leukemia in Ohio. Pediatr Blood Cancer 2007; 49(6):797-802.
71. Elwert F, Winship C. Endogenous selection bias: The problem of conditioning on a collider variable. Annual Review of Sociology 2014; 40:31-53.
72. Kinlen LJ, Hudson C. Childhood leukaemia and poliomyelitis in relation to military encampments in England and Wales in the period of national military service, 1950-63. BMJ 1991; 303(6814):1357-1362.
73. Dockerty JD, Cox B, Borman B, Sharples K. Population mixing and the incidence of childhood leukaemias: retrospective comparison in rural areas of New Zealand. BMJ 1996; 312(7040):1203-1204.
74. Kinlen LJ. Epidemiological evidence for an infective basis in childhood leukaemia. Br J Cancer 1995; 71(1):1-5.
75. Wainer H. The most dangerous equation. American Scientist 2007; 95(3):249-256.
76. Tu YK, Gilthorpe MS. The most dangerous hospital or the most dangerous equation? BMC Health Serv Res 2007; 7:185.
77. Kahneman D. Thinking, fast and slow. Macmillan; 2011.
78. Kinlen LJ. An examination, with a meta-analysis, of studies of childhood leukaemia in relation to population mixing. Br J Cancer 2012; 107(7):1163-1168.
79. Keijzer-Veen MG, Euser AM, van MN, Dekker FW, Vandenbroucke JP, van Houwelingen HC. A regression model with unexplained residuals was preferred in the analysis of the fetal origins of adult diseases hypothesis. J Clin Epidemiol 2005; 58(12):1320-1324.
80. Kline RB. Principles and Practice of Structural Equation Modeling. 3rd ed. New York: Guilford Press; 2015.
81. Hernan MA, Robins JM. Instruments for causal inference: an epidemiologist's dream? Epidemiology 2006; 17(4):360-372.
82. Wright S. The method of path coefficients. The annals of mathematical statistics 1934; 5(3):161-215.
83. Arnold KF, Gadd SC, Ellison GTH, Textor J, Gilthorpe MS. Adjusting for time-invariant and time-variant confounders in 'unexplained residuals' (UR) models. Journal of Epidemiology & Community Health 2016; 70(Supple 1):A61-A62.
84. Arnold KF, Gadd SC, Ellison GTH, Textor J, Tennant PWG, Heppenstall A et al. Challenges associated with 'unexplained residuals' (UR) models for longitudinal data within a causal framework, including adjustment for time-invariant and time-variant confounders. Stat Methods Med Res 2017; Submitted.
85. Tu YK, Gilthorpe MS. Unexplained residuals models are not solutions to statistical modeling of the fetal origins hypothesis. J Clin Epidemiol 2007; 60(3):318-319.
86. Andersen B. Methodological errors in medical research. London: Blackwell; 1990.
87. Moreno LF, Stratton HH, Newell JC, Feustel PJ. Mathematical coupling of data: correction of a common error for linear calculations. Journal of Applied Physiology 1986; 60:335-343.
88. Stratton HH, Feustel PJ, Newell JC. Regression of calculated variables in the presence of shared measurement error. Journal of Applied Physiology 1987; 62(5):2083-2093.

89. Tu YK, Gilthorpe MS, Griffiths GS. Is reduction of pocket probing depth correlated with the baseline value or is it "mathematical coupling"? J Dent Res 2002; 81(10):722-726.
90. Archie JP. Mathematical Coupling: A common source of error. Annals of Surgery 1981; 193:296-303.
91. Tu Y-K, Nelson-Moon ZL, Gilthorpe MS. Misuses of correlation and regression analyses in orthodontic research: the problem of mathematical coupling. American Journal of Orthodontics and Dentofacial Orthopedics 2004.
92. Tu YK, Maddick IH, Griffiths GS, Gilthorpe MS. Mathematical coupling can undermine the statistical assessment of clinical research: illustration from the treatment of guided tissue regeneration. J Dent 2004; 32(2):133-142.
93. Tu YK, Nelson-Moon ZL, Gilthorpe MS. Misuses of correlation and regression analyses in orthodontic research: the problem of mathematical coupling. Am J Orthod Dentofacial Orthop 2006; 130(1):62-68.
94. Altman DG. Statistics in medical journals. Statistics in Medicine 1982; 1(1):59-71.
95. Altman DG. Statistics in medical journals: developments in the 1980s. Statistics in Medicine 1991; 10(12):1897-1913.
96. Tu YK, Gilthorpe MS. Revisiting the relation between change and initial value: a review and evaluation. Stat Med 2007; 26(2):443-457.
97. Roberts JD, Fineman JR, Morin III FC. Inhaled nitric oxide and persistent pulmonary hypertension of the newborn. New England Journal Of Medicine 1997; 336:605-610.
98. Koh KK, Mincemoyer R, Bui MN, Csako G, Pucino F, Guetta V et al. Effects of hormone-replacement therapy on fibrinolysis in postmenopausal women. New England Journal Of Medicine 1997; 336:683-690.
99. Fleury S, Rizzardi GP, Chapuis A, Tambussi G, Knabenhans C, Simeoni E et al. Long-term kinetics of T cell production in HIV-infected subjects treated with highly active antiretroviral therapy. Proceedings of the National Academy of Sciences of the United States of America 2000; 97(10):5393-5398.
100. Marcus R, Holloway L, Wells B, Greendale G, James MK, Wasilauskas C et al. The relationship of biochemical markers of bone turnover to bone density changes in postmenopausal women: results from the postmenopausal estrogen/ progestin interventions (PEPI) trial. Journal of Bone and Mineral Research 2001; 14:1583-1595.
101. Uchino K, Billheimer D, Cramer SC. Entry criteria and baseline characteristics predict outcome in acute stroke trials. Stroke 2001; 32:909-916.
102. Diverse Populations Collaborative Group. Weight-height relationships and body mass index: some observations from the Diverse Populations Collaboration. Am J Phys Anthropol 2005; 128(1):220-229.
103. Oldham PD. A note on the analysis of repeated measurements of the same subjects. Journal of Chronic Diseases 1962; 15:969-977.
104. Altman DG. Practical statistics for Medical Research. London: Chapman and Hall/CRC; 1991.
105. Tu Y-K, Gilthorpe MS. Statistical Thinking. New York: Chapman & Hall; 2012.
106. Pitman E. A Note on Normal Correlation. Biometrika 1939; 31:9-12.
107. Morgan W. A test for the significance of the differences between two variances in a sample from a normal bivariate distribution. Biometrika 1939; 31:13-19.
108. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986; 1:307-310.
109. Gilthorpe MS, Maddick IH, Petrie A. Introduction to multilevel modelling in dental research. Community Dental Health 2000; 17(4):222-226.
110. Gilthorpe MS, Griffiths GS, Maddick IH, Zamzuri AT. The application of multilevel modelling to periodontal research data. Community Dental Health 2000; 17(4):227-235.
111. Blance A, Tu YK, Gilthorpe MS. A multilevel modelling solution to mathematical coupling. Stat Methods Med Res 2005; 14(6):553-565.
112. Pearson K. On a form of spurious correlation which may arise when indices are used in the measurement of organs. Proceedings of the Royal Society of London 1897; 60:489-498.
113. Tu YK, Law GR, Ellison GT, Gilthorpe MS. Ratio index variables or ANCOVA? Fisher's cats revisited. Pharm Stat 2009.
114. Rubin DB. Causal inference using potential outcomes: design, modeling, decisions. American Statistical Association 2007; 100:322-331.
115. Rubin DB, Waterman RP. Estimating the causal effects of marketing interventions using propensity score methodology. Statistical Science 2006; 21:206-222.
116. Ridout M, Demétrio CGB, Hinde J. Models for count data with many zeros. Proceedings of the XIXth International Biometric Conference 1998;179-192.
117. Mullahy J. Specification and Testing of Some Modified Count Data Models. Journal of Econometrics 1986; 33(3):341-365.

118. Lambert D. Zero-Inflated Poisson Regression, with An Application to Defects in Manufacturing. Technometrics 1992; 34(1):1-14.
119. Böhning D. Zero-Inflated Poisson Models and C.A.MAN: A Tutorial Collection of Evidence. Biometrical Journal 1998; 40(7):833-843.
120. Gilthorpe MS, Frydenberg M, Cheng Y, Baelum V. Modelling count data with excessive zeros: the need for class prediction in zero-inflated models and the issue of data generation in choosing between zero-inflated and generic mixture models for dental caries data. Stat Med 2009; 28(28):3539-3553.
121. Hall DB. Zero-Inflated Poisson and Binomial Regression with Random Effects: A Case Study. Biometrics 2000; 56(4):1030-1039.
122. Vieira AMC, Hinde JP, Demetrio CGB. Zero-inflated proportion data models applied to a biological control assay. Journal of Applied Statistics 2000; 27(3):373-389.
123. Böhning D, Dietz E, Schlattmann P. The zero-inflated Poisson model and the decayed, missing and filled teeth index in dental epidemiology. Journal of the Royal Statistical Society, Series A 1999; 162:195-209.
124. Lord FM. A paradox in the interpretation of group comparisons. Psychological Bulletin 1967; 68:304-305.
125. Lord FM. Statistical adjustments when comparing preexisting groups. Psychological Bulletin 1969; 72:337-338.
126. Blance A, Tu YK, Baelum V, Gilthorpe MS. Statistical issues on the analysis of change in follow-up studies in dental research. Community Dentistry and Oral Epidemiology 2007; 35(6):412-420.
127. Skrondal A, Rabe-Hesketh S. Generalized Latent Variable Modeling: Multilevel, Longitudinal and Structural Equation Models. London: Chapman & Hall; 2004.
128. Vermunt JK, Magidson J. Technical Guide for Latent GOLD 4.0: Basic and Advanced. Massachusetts: Statistical Innovations Inc.; 2005.
129. Bland JM, Altman DG. Measuring agreement in method comparison studies. Statistical Methods in Medical Research 1999; 8(2):135-160.
130. Wong MC, Schwarz E, Lo EC. Patterns of dental caries severity in Chinese kindergarten children. Community Dentistry and Oral Epidemiology 1997; 25(5):343-347.
131. Holst D. The relationship between prevalence and incidence of dental caries. Some observational consequences. Community Dental Health 2006; 23(4):203-208.
132. Groeneveld A. Longitudinal study of prevalence of enamel lesions in a fluoridated and non-fluoridated area. Community Dentistry and Oral Epidemiology 1985; 13(3):159-163.
133. Poulsen S, Heidmann J, Vaeth M. Lorenz curves and their use in describing the distribution of 'the total burden' of dental caries in a population. Community Dental Health 2001; 18(2):68-71.
134. Smith ADAC, 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.
135. 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. Int J Epidemiol 2016; 45(4):1271-1279.
136. Goldstein H. The multilevel analysis of growth data. In: Hauspie R, Lindren G, Falkner F, editors. Essays on Auxology, presented to JM Tanner. Castlemead; 1996. 39-52.
137. Gilthorpe MS, Zamzuri AT, Griffiths GS, Maddick IH, Eaton KA, Johnson NW. Unification of the "burst" and "linear" theories of periodontal disease progression: a multilevel manifestation of the same phenomenon. J Dent Res 2003; 82(3):200-205.
138. Bollen K, Curran P. Latent curve models. 2nd ed. New York: Wiley; 2006.
139. Duncan TE, Duncan SE, Stryker LA. An introduction to latent variable growth curve modeling. 2nd ed. Mahwah, NJ: Laurence Erlbaum Associates Inc; 2011.
140. Pickles A, Croudace T. Latent mixture models for multivariate and longitudinal outcomes. Stat Methods Med Res 2010; 19(3):271-289.
141. Kubzansky LD, Gilthorpe MS, Goodman E. A prospective study of psychological distress and weight status in adolescents/young adults. Ann Behav Med 2012; 43(2):219-228.
142. Gong Y, Hounsa A, Egal S, Turner PC, Sutcliffe AE, Hall AJ et al. Postweaning exposure to aflatoxin results in impaired child growth: a longitudinal study in Benin, West Africa. Environ Health Perspect 2004; 112(13):1334-1338.
143. Goldstein H. Multilevel Statistical Models. London: Edward Arnold; 1995.
144. Bryk AS, Raudenbush AW. Hierarchical Linear Models: Applications and Data Analysis Methods. London: Sage; 1992.
145. Curran PJ. Have Multilevel Models Been Structural Equation Models All Along? Multivariate Behavioral Research 2003; 38(4):529-569.

146. Jung T, Wickrama KAS. An introduction to latent class growth analysis and growth mixture modeling. Social and Personality Psychology Compass 2008; 2:302-317.
147. Nagin DS, Odgers CL. Group-based trajectory modeling in clinical research. Annu Rev Clin Psychol 2010; 6:109-138.
148. Nagin DS. Analyzing developmental trajectories: A semiparametric, group-based approach. Psychological Methods 1999; 4(2):139-157.
149. Gilthorpe MS, Dahly DL, Tu YK, Kubzansky LD, Goodman E. Challenges in modelling the random structure correctly in growth mixture models and the impact this has on model mixtures. Journal of Developmental Origins of Health and Disease 2014;1-9.
150. Aflatoxin: scientific background, control, and implications. Elsevier; 2012.
151. Leong YH, Latiff AA, Ahmad NI, Rosma A. Exposure measurement of aflatoxins and aflatoxin metabolites in human body fluids. A short review. Mycotoxin Res 2012; 28(2):79-87.
152. Khlangwiset P, Shephard GS, Wu F. Aflatoxins and growth impairment: a review. Crit Rev Toxicol 2011; 41(9):740-755.
153. Turner PC. The molecular epidemiology of chronic aflatoxin driven impaired child growth. Scientifica (Cairo) 2013; 2013:152879.
154. Tu Y-K, Gilthorpe MS. Univariate and multivariate data analysis. In: Wild CP, Vineis P, Garte S, editors. Molecular Epidemiology of Chronic Disease. London: Wiley; 2008.
155. World Health Organisation. The International Statistical Classification of Diseases and Health Related Problems ICD-10: Tenth Revision. 2nd ed. Geneva: World Health Organisation; 2005.
156. Dukes CE. The surgical pathology of rectal cancer. J Clin Pathol 1949; 2(2):95-98.
157. Townsend P, Phillimore P, Beattie A. Health and Deprivation: Inequality and the North. London: Routledge; 1988.
158. Schrijvers CT, Mackenbach JP, Lutz JM, Quinn MJ, Coleman MP. Deprivation, stage at diagnosis and cancer survival. Int J Cancer 1995; 63(3):324-329.
159. Pollock AM, Vickers N. Breast, lung and colorectal cancer incidence and survival in South Thames Region, 1987-1992: the effect of social deprivation. J Public Health Med 1997; 19(3):288-294.
160. Wrigley H, Roderick P, George S, Smith J, Mullee M, Goddard J. Inequalities in survival from colorectal cancer: a comparison of the impact of deprivation, treatment, and host factors on observed and cause specific survival. J Epidemiol Community Health 2003; 57(4):301-309.
161. Lyratzopoulos G, Sheridan GF, Michie HR, McElduff P, Hobbiss JH. Absence of socioeconomic variation in survival from colorectal cancer in patients receiving surgical treatment in one health district: cohort study. Colorectal Dis 2004; 6(6):512-517.
162. Fuller WA. Measurement error models. New York: Wiley; 1987.
163. Greenwood DC, Gilthorpe MS, Cade JE. The impact of imprecisely measured covariates on estimating gene-environment interactions. BMC Med Res Methodol 2006; 6:21.
164. Quirke P, Morris E. Reporting colorectal cancer. Histopathology 2007; 50(1):103-112.
165. Morris EJ, Maughan NJ, Forman D, Quirke P. Identifying stage III colorectal cancer patients: the influence of the patient, surgeon, and pathologist. J Clin Oncol 2007; 25(18):2573-2579.
166. Morris EJ, Maughan NJ, Forman D, Quirke P. Who to treat with adjuvant therapy in Dukes B/stage II colorectal cancer? The need for high quality pathology. Gut 2007; 56(10):1419-1425.
167. Niggebrugge A, Haynes R, Jones A, Lovett A, Harvey I. The index of multiple deprivation 2000 access domain: a useful indicator for public health? Soc Sci Med 2005; 60(12):2743-2753.
168. Robinson WS. Ecological Correlations and the Behavior of Individuals. American Sociological Review 1950; 15(3):351-357.
169. VanderWeele TJ. Explaination in Causal Inference: Methods for Mediation and Interaction. New York: Oxford University Prress; 2015.
170. Walter SD, Holford TR. Additive, multiplicative, and other models for disease risks. Am J Epidemiol 1978; 108(5):341-346.
171. Kupper LL, Hogan MD. Interaction in epidemiologic studies. Am J Epidemiol 1978; 108(6):447-453.
172. Saracci R. Interaction and synergism. Am J Epidemiol 1980; 112(4):465-466.
173. Rothman KJ, Greenland S, Walker AM. Concepts of interaction. Am J Epidemiol 1980; 112(4):467-470.
174. Rowntree RK, Harris A. The phenotypic consequences of CFTR mutations. Ann Hum Genet 2003; 67(Pt 5):471-485.
175. Vandenbroucke JP, Koster T, Briët E, Reitsma PH, Bertina RM, Rosendaal FR. Increased risk of venous thrombosis in oral-contraceptive users who are carriers of factor V Leiden mutation. Lancet 1994; 344:1453-7.
176. Vandenbroucke JP, van der Meer FJM, Helmerhorst FM, Rosendaal FR. Factor V Leiden: should we screen oral contraceptive users and pregnant women? British Medical Journal 1996; 313:1127-30.

177. Ganzach Y. Misleading interaction and curvilinear terms. Psychological Methods 1997; 2(3):235-47.
178. Greenland S. Tests for interaction in epidemiologic studies: a review and a study of power. Stat Med 1983; 2(2):243-251.
179. Rees DC, Cox M, Clegg JB. World distribution of factor V Leiden. Lancet 1995; 346(8983):1133-1134.
180. Clayton D, McKeigue PM. Epidemiological methods for studying genes and environmental factors in complex diseases. Lancet 2001; 358(9290):1356-1360.
181. Thompson WD. Effect modification and the limits of biological inference from epidemiologic data. J Clin Epidemiol 1991; 44(3):221-232.

Advanced Modelling Strategies: Challenges and pitfalls in robust causal inference with observational data summarises the lecture notes prepared for a four-day workshop sponsored by the *Society for Social Medicine* and hosted by the *Leeds Institute for Data Analytics (LIDA)* at the University of Leeds on 17th-20th July 2017.

ISBN 978-1-5272-1208-4

LEEDS Institute for Data Analytics