Short running head
Concerning Replicability

Full title
Statistical and Scientific Considerations Concerning the Interpretation, Replicability and
Transportability of Research Findings

Authors
Richard J. Cook, Department of Statistics and Actuarial Science, University of Waterloo
ORCID ID: 0000-0002-1414-4908
Jerald F. Lawless, Department of Statistics and Actuarial Science, University of Waterloo
ORCID ID: 0000-0002-1414-4908

Sources of Support
This work was funded by the Natural Sciences and Engineering Research Council of Canada through
Discovery Grants to RJC (RGPIN-2017-04207) and JFL (RGPIN-2017-04055).

Conflict of Interest
The authors declare no potential conflict of interests.

Corresponding author contact information
Richard J. Cook Department of Statistics and Actuarial Science, University of Waterloo, 200
University Avenue West, Waterloo, Ontario N2L 3G1 email: rjcook@uwaterloo.ca
Abstract

To advance scientific understanding of disease processes and related intervention effects, study results should be free from bias and replicable. More broadly, investigators seek results that are transportable, meaning applicable to a perceived study population and also in other environments and populations. We review fundamental statistical issues which arise in the analysis of observational data from disease cohorts and other sources and discuss how these issues affect the transportability and replicability of research results. Much of the literature focuses on estimating average exposure or intervention effects at the population level, but we argue for more nuanced analyses of conditional effects that reflect the complexity of disease processes.

Key Words: Cohort studies, databases, outcomes, prognosis, registries, risk factors.
1. Introduction

Replicability is foundational to scientific progress in the understanding and treatment of disease processes. Benjamini\(^1\) quotes the geneticist and pioneer of statistical science R.A. Fisher as saying that “no isolated experiment, however significant by itself, can suffice for the experimental demonstration of any natural phenomenon”, and goes on to emphasize “replicated discovery”.\(^2\) Popper\(^3\) likewise states that “non-replicable single occurrences are of no significance to science”. While the concept of replicability was originally considered in connection with experimental studies, it has more recently been discussed in connection with observational studies and analyses involving so-called “real world data” (RWD).

There has been considerable discussion of so-called replicability crises in medical science related to randomized trials (e.g., Ioannidis\(^4\), Begley and Ellis\(^5\)), controlled laboratory experiments\(^6\), and observational studies\(^7,8\). These discussions arose from the fact that findings from many (some would say most) published studies cannot be replicated by other researchers. We consider reasons why study findings may not be replicated in future studies, and why results and conclusions may not be transportable. Benjamini\(^1\) has discussed two related reasons; one is the use of invalid measures of uncertainty concerning study results and the other is the use of "selective" inference in which large numbers of potential effects are examined (e.g., through hypothesis tests) and only the strongest (e.g., highly significant) ones are identified and reported. Our focus is on issues related to the ways studies are designed, run and analyzed; they include the selection of study participants and the collection and analysis of data about them. Some important aspects are subtle, especially in observational studies, and studies or analyses that fail to address them adequately can produce biased inferences.

We focus here on estimation, testing and replication of what we term “effects”. This includes the effects of an intervention in an experimental study, the effect of an exposure (or other risk factor)
on the occurrence of a health-related event in an observational study, the effect of genetic factors on health outcomes, and the effect of prior health conditions or outcomes on potential future events. Replication of findings concerning an individual effect means that estimates of the effect in an original study and a subsequent study are similar, given the margins of error involved in each estimate. We emphasize that effects can be defined in different ways, depending on the nature of the outcome and on assumed conditions and models, and that the clear specification of effects is a clinical as well as statistical matter.

The remainder of the paper is as follows. In Section 2 we introduce some terminology associated with research studies. In Section 3 we formally define conditional and average effects and introduce models which accommodate heterogeneity of effects. In Section 4 we review concepts of causation and consider the potential roles of confounding, selection, and other factors in biasing inferences about covariate or treatment effects. Extensions to deal with time-varying processes are discussed briefly in Section 5 and Section 6 discusses reasons why results concerning effects and other features of a process may not be consistent across studies, and ways that studies can be designed and analysed to enhance replicability and transportability. Section 7 summarizes the discussion and makes some general recommendations for disease studies. Throughout we illustrate concepts and methods with references to studies in psoriatic arthritis and other rheumatic diseases. A companion follow-up paper will consider observational studies of time-varying processes in more detail and discuss the estimation and understanding of intervention effects.

2. Some Terminology Regarding Research Findings

The notion of replication is one of R.A. Fisher’s many contributions to scientific method. Within an experiment, replicates are formed when independent experimental units receive the same treatment in the same fashion as other units; the process of administering treatments in this fashion (i.e., creating replicates) is called replication. When assessing whether similar findings are obtained
over multiple studies in a body of scientific research, replication has a different meaning. In health research a study finding is *replicable* if other studies with similar ways of selecting participants, managing care, collecting data, and conducting analyses tend to produce a similar result. The terms “tend to” and “similar” recognize the fact that environmental and temporal factors inevitably vary across studies, and that in finite samples there is always sampling variation. Thus, results are not guaranteed to be close to identical even when study populations and protocols in two studies are similar. As noted in the introduction, we focus here on estimation and replicability of effects associated with interventions, exposures and other factors affecting disease processes. Effects are defined formally in Section 3.

Replicability is related to the less formal but broader concepts of generalizability and transportability. *Generalizability* refers to how well a finding in a study applies to a broader target population and is often used when assessing the interpretation and relevance of findings from clinical trials to typical patients. *Transportability* refers to how well study findings apply in a different (but not necessarily broader) population.\(^{10,11}\) Both terms are related to concepts of validity: *internal validity* of study results is often taken to mean the generalizability (or transportability) of results to a specified study population, while *external validity* refers to transportability of results to other target populations. This meaning of validity is in the spirit of Keiding and Louis.\(^{11}\) Some authors define validity more narrowly, for example to refer to unbiased estimation of an effect in some population; see Degtiar and Rose.\(^{12}\) We will use it here in the former sense but the transportability in question often concerns some effect, and so absence of bias in an estimator of the effect is an important aspect.

Findings may be nonreplicable or nontransportable for biological reasons, but it may also be due to sampling, selection, and analytical biases – this is a central theme of the paper. The terms replicability, generalizability and transportability should be distinguished from the term reproducibility, which has sometimes been used in the same sense as replicability.\(^1\) Study findings
are said to be *reproducible* if the study data, analysis plan and code used for the analysis are provided in sufficient detail that someone else can reproduce the findings and corroborate the conclusions of the study. This is the implicit goal in the modern movement for reproducible research.\textsuperscript{13}

Disease processes are often complex, and multiple aspects or features may be of interest in a study. At the planning stage, good studies should specify key features of interest and the primary targets for inference. Even in randomized controlled trials (RCTs) this can be challenging. For example, in a trial investigating a new medical treatment versus standard care for persons with heart disease, there may be several events of interest, including myocardial infarction, stroke, hospitalization, death from cardiovascular causes, or death from other causes. Moreover, individuals in a trial may require other medical or surgical interventions which may complicate comparisons between treatment arms. Bührer et al.\textsuperscript{14} discuss aspects of RCTs that bear on the difficulty of defining estimands and replicating a trial and its results.

Morand et al.\textsuperscript{15} describe a phase 3 trial involving a human monoclonal antibody (anifrolumab) for persons with systemic lupus erythematosus (SLE). A phase 2 trial had shown a reduction in disease activity measured by the responder index-4 score SRI-4 but in the phase 3 trial there was no significant difference in this score, measured 52 weeks after randomization, in the treatment and placebo control groups. There was, however, a significant effect on a secondary response, namely the BICLA composite lupus assessment score after 52 weeks. Here we have an example of an effect for one outcome replicating, but not the effect for another outcome. Another smaller study by Furie et al.\textsuperscript{16} likewise found no effect on the primary response; no significant effects on the secondary responses were observed either, but the directions of effects were similar to those of Morand et al.\textsuperscript{15}

3. Conditional and Average Effects
Clinical trials aim to estimate the causal effect of an experimental treatment on a clinically important outcome; often the goal is to alleviate symptoms more effectively, or to slow or prevent the progression of joint damage. Inclusion and exclusion criteria are defined to specify the study population of interest and typically analyses are directed at assessing whether the average response in the treated arm is better than the average response in the control arm. For example, Keystone et al.\textsuperscript{17} report on a double-blind multicenter trial examining the effect of 40 mg or 20 mg of adalimumab administered subcutaneously every other week versus a placebo injection on progression of structural joint damage at 52 weeks. The authors found a highly significant ($p < 0.001$) reduction in the progression of joint damage as measured by the total Sharp Score (tSS); for those in the arm receiving 40 mg of adalimumab, the mean increase in tSS over 52 weeks was 0.1 while those in the placebo arm it was 2.7. This gives a difference of $-2.6$ for adalimumab versus placebo; we refer to this difference in the average reduction of tSS as an average treatment effect, as defined below.

Jamal et al.\textsuperscript{18} considered secondary analyses of this same trial with the aim of assessing the effect-modifying role of various patient characteristics including disease duration ($\leq 3$ years corresponding to early disease versus $> 3$ years for those with later more established disease). A significant duration-treatment interaction was found such that patients receiving 40 mg of adalimumab with early disease had a 5.32 lower increase in the tSS over 52 weeks, compared to those on placebo, whereas in patients with established disease this reduction was only 2.06 ($p = 0.048$). Such analyses provide insights into what kinds of patients may benefit more or less from treatment, and the resulting treatment effects are called conditional effects – they correspond to patients satisfying particular conditions.

We now introduce notation and models to discuss effects in empirical studies more precisely. Letters are conventionally used to denote features measured on individuals, for example, genetic, demographic or clinical traits, as well as interventions and outcomes. We begin by considering a
simple context where there is a binary experimental treatment (or exposure) variable $X$ with $X = 1$ if an individual is treated or exposed, and $X = 0$ otherwise. The outcome or response variable is denoted by $Y$ and a collection of $p$ auxiliary covariates is denoted by $W = (W_1, \ldots, W_p)'$; the bold font is used when one letter represents multiple (here $p$) covariates. The features $Y$, $X$, and $W$ are called variables because they *vary* across individuals in a population. In the context of the studies by Keystone et al.\textsuperscript{17} and Jamal et al.\textsuperscript{18} the response $Y$ is the change in the tSS over 52 weeks, $X = 1$ indicates adalimumab treatment, $X = 0$ indicates placebo, and $W$ is a single variable that indicates early disease.

Probability distributions are used to characterize variation in a population; the distribution for the full set of variables is denoted $P(Y, X, W)$ with $P(Y \mid X = x, W = w)$ denoting the “conditional” distribution of $Y$ for individuals with a specific set of values for $X$ and $W$ represented by the lower case letters $x$ and $w = (w_1, \ldots, w_p)'$. Such distributions are crucial for summarizing variation and predicting patient outcomes, but we will focus here on their role in defining effects. Specifically, such models can be used to describe the effect of an exposure $X$ or set of covariates $W$ on the distribution of $Y$. For simplicity we consider the case where $Y$ is a continuous response. We focus on linear models characterizing how the expected value (average, or mean) of $Y$ varies according to $X$ and $W$, or possibly just $X$. For the former, $E(Y \mid X = x, W = w)$ denotes the mean of $Y$ among individuals with specific values of $X$ and $W$. Linear models involve the specification

$$E(Y \mid X = x, W = w) = \beta_0 + \beta_1 x + \beta_2 w$$

(3.1)

where $\beta_2 w = \beta_{21} w_1 + \cdots + \beta_{2p} w_p$ is a compact way of writing the linear combination of the $\beta_{2j} W_j$ terms, $j = 1, \ldots, p$. We refer to the parameter $\beta_1$ in (3.1) as a *conditional effect* of $X$ on $Y$ because we condition on $W = w$; it represents the difference in the mean of $Y$ for two persons with $X = 1$ and $X = 0$ respectively, but with the same values for covariates $W$. 


A model that accommodates heterogeneity (or modification, according to the value of \( W \)) of the treatment effect is obtained by adding an \( X \cdot W \) interaction term to (3.1) to give

\[
E(Y|X = x, W = w) = \beta_0 + \beta_1 x + \beta_2' w + \beta_3' x w. \tag{3.2}
\]

This leads to conditional effects given by

\[
E(Y|X = 1, W = w) - E(Y|X = 0, W = w) = \beta_1 + \beta_3' w, \tag{3.3}
\]

which vary according to the \( W \)-values held by an individual. Such a model was fitted by Jamal et al.\(^{18}\) when they identified the modifying role of disease duration in the effect of adalimumab on the 52 week tSS. The treatment effect is said to be heterogeneous when \( \beta_3 \neq 0 \) and homogeneous (i.e. the same across the members of the population) when \( \beta_3 = 0 \), in which case (3.1) applies.

The marginal (or average) effect of \( X \) quantifies the difference between individuals with \( X = 1 \) and \( X = 0 \), averaged over the auxiliary covariates \( W \). From (3.2) this equals

\[
E(Y|X = 1) - E(Y|X = 0) = \beta_1 + \beta_3' E(W|X = 1) - E(W|X = 0) + \beta_3' E(W|X = 1). \tag{3.4}
\]

We notice that the average effect of \( X \) equals \( \beta_1 \) if \( i) \) and \( W \) are independent (denoted by \( X \perp W \)) so that \( E(W|X = 1) = E(W|X = 0) = E(W) \), and \( ii) \) there is no \( X \cdot W \) interaction (i.e. \( \beta_3 = 0 \)). In all other settings the average effect depends on the distribution of \( W \) given \( X \) and even in randomized experiments, it depends on \( E(W) \). Thus, in the presence of heterogeneity, the average treatment effect is relevant to the study population represented in the trial, but this may have little bearing on a patient being considered for treatment in a target population. The principle of stratified medicine would suggest reporting the effect of adalimumab on 52 week reduction in tSS according to patient disease duration.

Conditional effects based on comprehensive and physically plausible models that represent observed data adequately are more likely to be transportable. Since estimates of marginal effects are obtained by averaging conditional effects over values of \( W \), average effects are less transportable to populations with different distributions of \( W \); see Keiding and Louis.\(^{11}\)
Interestingly, average treatment effects tend to be favoured in randomized trials because of their simple causal interpretation, the fact that they appear to rely on minimal assumptions, and that they can be viewed as internally valid, but the same issue affects the replicability and transportability of results from randomized trials. If a new trial involves a study population with a different distribution of prognostic factors $W$, average treatment effects will be difficult to replicate. Finally, the transportability of study results even to a perceived study population ("internal" validity) may be compromised by various factors, including study selection effects (e.g., refusal of potential subjects to participate), inadequate handling of confounders, measurement error, missing data, and disease-related loss to follow-up.

We now turn to a discussion of causal effects.

4. Causal Effects, Confounding and Selection Bias

4.1 Causal Effects

Science seeks to understand causal mechanisms, but Dawid\textsuperscript{19} characterizes causality as a "slippery and ambiguous concept". Frameworks for causal reasoning have received a great deal of attention in recent years, especially with respect to the effects of exposures or interventions in epidemiology and medicine. Different schools of thought have emerged; see for example Dawid,\textsuperscript{20} Arjas and Parner,\textsuperscript{21} Rubin\textsuperscript{22} along with associated discussions. Moodie and Stephens\textsuperscript{23} give a thoughtful recent review. Here we take a broad view wherein careful collection and analysis of data from specific studies, supported by background scientific knowledge, may lead us to label (with caution) certain effects as causal.

Causal effects of a binary exposure or intervention are frequently defined through the conceptualization of potential outcomes which would occur under treatments $X = 0$ and $X = 1$ respectively while all other conditions are held fixed. In this framework one imagines outcomes
Y(0) and Y(1) as pre-existing for a given individual, with Y(0) or Y(1) revealed through observation if
X = 0 or X = 1 respectively. Such a conceptualization is not realistic in actual studies where an
individual can receive only one treatment, but a detailed and formal theory of causality has
nevertheless been based on such potential outcomes.

The random assignment of treatment to individuals in an experimental study provides a basis
for causal inference concerning treatment effects. However, in an observational study the exposure
or treatment for a given individual is often associated with auxiliary factors W that are also related
to the response Y; such variables W are referred to as confounders. More formally, W is a
confounding variable for the X–Y association if it satisfies two conditions. First W must be
associated with the exposure for reasons other than a causal effect of the exposure on W. Second, a
W–Y association must exist due to W being a cause or a correlate of a cause for the outcome, or due
to W being associated with the identification of the outcome. If all such factors are known and
appropriately accounted for in a model such as (3.1) or (3.2) then we say there are no hidden or
unmeasured confounders, and the conditional effect of X on Y given by β_1 in (3.3) may have a causal
interpretation. In what follows when we refer to conditional or average effects as having a causal
interpretation, we assume implicitly that such unverifiable assumptions are satisfied – the evidence
for a causal effect depends on the scientific plausibility of such assumptions. If some confounding
variables in W are unknown or unaccounted for, no such inference can be made. Model checking is
possible, but one can rarely be confident that there are no unknown confounders (there are
invariably “unknown unknowns”). Dawid\textsuperscript{19} remarks that assumptions such as this needed to
support causal inference “should never be accepted glibly or automatically, but deserve careful
attention and context-specific discussion and justification whenever the methods are applied.”

In an experimental study, random allocation of individuals to treatment ensures that X is
independent of all covariates, including unobserved confounders. In that case it may be argued that
β_1 in (3.1) is a conditional causal effect. The average causal effect (ACE) is defined by averaging (3.3)
or (3.4) over values of \( W \). Since \( W \) and \( X \) are independent here, they both give \( ACE = \beta_1 + \beta_2 E(W) \) from which it can be seen that the \( X-W \) interaction in (3.2) makes the average causal treatment effect depend on the distribution of \( W \). In settings with \( X-W \) interactions, an \( ACE \) obtained by directly fitting a linear model for \( E(Y|X) \) is of limited relevance since different patients (or types of patients) benefit from a treatment to different degrees. In clinical trials, the knowledge that treatment effects may be heterogeneous motivates proper subgroup analyses.\(^{25}\) In multi-center trials, for example, treatment by center interactions are often seen when the mix of patients and standard of care varies across centers. Indeed, the existence of interactions motivates research into personalized and stratified medicine.

For observational studies where \( W \) and \( X \) cannot be expected to be independent, a variety of methods for obtaining an \( ACE \) estimate (under stringent assumptions) have been developed. An \( ACE \) in the present context is defined as the marginal treatment effect that would be obtained if the conditional mean \( E(Y|X, W) \) is the same as in the study population, but that \( W \) is independent of \( X \), with a specified distribution. This can be thought of as emulating a hypothetical experimental study in which \( X \) was set by randomization.\(^{24}\) The distribution for \( W \) is conventionally taken to be its marginal distribution in the study population. Methods for estimating an \( ACE \) include regression, wherein a model for \( Y \) given \( X \) and \( W \) is fitted and then an \( ACE \) is estimated by averaging the fitted means for each exposure value \((X = 0, 1)\) with respect to \( W \), under the assumption that the distribution of \( W \) is independent of \( X \) and the same as in the population of interest. Another common approach is to utilize so-called propensity score weighting, which requires fitting a model for \( P(X|W) \), where \( P(X = 1|W) \) is called the propensity to treat. A third approach is instrumental variable regression – see Angrist et al.\(^{27}\) All three methods rely on assumptions that cannot be checked using just the observed data. Moodie and Stephens\(^{21}\) give references to these and other methods and provide a thoughtful discussion of the assumptions needed. They include the assumption that there are no unmeasured confounders and the positivity condition \( P(X = 1|W) > 0 \) for all \( W \); the latter
means that for every individual in the population (i.e., for all values of $W = w$) there is a non-zero chance of receiving treatment (or being exposed). One or both of these assumptions are typically violated in observational studies.

It is important, especially in observational settings, to think carefully about relationships between variables. Directed acyclic graphs\textsuperscript{26}, or DAGs for short, are graphical representations of the relations between variables that can facilitate discussion of causal analysis.\textsuperscript{24,29} DAGs are comprised of nodes representing variables, with arrows used to represent a causal effect from one variable to another; that is, a change in the variable from which the arrow emanates leads to a change in the variable to which the arrow points. Dawid\textsuperscript{20} uses the term “causal DAG” for this interpretation but notes that DAGs can also be viewed as representing dependence structures (i.e., conditional independencies). We adopt this latter view in what follows.

Figure 1(a) depicts an idealization of a setting in an arthritis clinic where a patient presents with features summarized in $W$, which influences the decision to prescribe a treatment and may also have a causal effect on the outcome $Y$. The variable $W$ might contain information on the erythrocyte sedimentation rate (ESR), for example, where high values tend to be associated with joint pain and swelling, which in turn may increase the probability that a biologic treatment is prescribed. Elevated ESR is also associated with progression of joint damage over the next six months, which may be the response $Y$ in a study. Controlling for $W$ when assessing the effect of $X$ on $Y$ (i.e., considering $E(Y \mid X, W)$) ensures that the treatment is being evaluated among individuals who are similar with respect to the features in $W$. If there are no hidden confounders and the model (3.2) is valid, this provides a conditional causal effect (CCE) estimate based on (3.3) which is transportable to populations with different distributions of $W$. It should be stressed, however, that conditional models should not be assumed to be internally valid without carefully checking underlying assumptions against observed data; model misspecification is a common source of bias in estimating conditional effects. An average causal treatment effect can also be obtained using one of
the methods mentioned above. Note that an average causal effect (ACE) is an “internal” effect in the sense that it applies to the study population. Replicability and transportability to other populations requires independent confirmation through comparison with external studies and data; our remarks are to indicate that conditional effect estimates are more likely to be transportable than average effects.

4.2 Confounding, Selection Effects and Collider Bias

Failure to deal with confounders in the analysis of observational studies is a common source of bias, particularly for average treatment effects. We now discuss confounders from the perspective of causal graphs (or DAGs), as well as another phenomenon called collider bias, which can arise from unaddressed selection bias and other sources and compromises estimation of both conditional and average effects. The DAG in Figure 1(a) represents a setting where there is a causal effect of $X$ on $Y$, conveyed by the arrow from $X$ to $Y$. However, the covariates $W$ include confounders and their omission in an analysis biases estimation of causal effects. As an extreme example, consider the DAG of Figure 2(a). Since there is no direct arrow from $X$ to $Y$ here, there is no causal effect of $X$ on $Y$. The symbol $\perp$ represents independence and we write $X \perp Y | W$ to indicate that there is no association between $X$ and $Y$ given $W = w$. Omission of $W$ in an analysis will, however, lead to a spurious association which may be erroneously interpreted as evidence of a causal effect; that is while $X \perp Y | W$, variables $X$ and $Y$ are not marginally independent. The nature of the apparent effect of $X$ on $Y$ depends on the effects of $W$ on both $X$ and $Y$. With a scalar confounder $W$, if an increase in $W$ causes an increase in both the mean of $X$ and $Y$, then it will appear that an increase in $X$ will yield an increase in the mean of $Y$. If an increase of $W$ causes an increase in the mean of $X$ and a decrease in the mean of $Y$, then treatment ($X = 1$) will appear to cause a reduction in the mean of $Y$. Both of these associations are entirely spurious.
Careful thought about the association between variables, and potential causal relations, is important. Confounding is a well-known phenomenon of which most researchers are aware, but selection biases are sometimes more subtle. These arise when individuals in a study are not representative of the intended study population. Figure 2(b) shows a DAG that once again involves the three variables $W$, $X$ and $Y$; here $X$ and $Y$ each have a causal effect on $W$, but since there is no arrow between $X$ and $Y$, again there is no causal link between them. A spurious association between $X$ and $Y$ is induced here, however, if we condition on $W$, through what is called a collider bias; the term reflects the feature of the DAG wherein two arrows point to (or collide at) $W$. Thus, in Figure 2(a) conditioning on $W$ reveals that there is no causal effect of $X$ on $Y$, while in Figure 2(b) conditioning on $W$ spuriously suggested a causal effect of $X$ on $Y$.

From the probabilistic interpretation of DAGs, Figure 2(a) conveys a conditional independence $X \perp Y | W$ where $P(Y | X, W) = P(Y | W)$. In Figure 2(b) however, while $X \perp Y \text{ and } P(Y | X) = P(Y)$, conditioning on $W$ destroys this independence: $X \nmid Y | W$. This is explained by the relation

$$P(Y | X, W) = \frac{P(Y | X)P(W | Y, X)}{P(W | X)}$$

Because there is an arrow from $Y$ to $W$, $P(W | Y, X) \neq P(W | X)$ and hence $P(Y | X, W) \neq P(Y | X) = P(Y)$.

Collider bias has received a great deal of attention in epidemiology in recent years, but we generally favour the more descriptive broad term selection bias, since the settings where collider bias arise typically involve analyses based on selective sub-samples of a study sample or study population. Biases can arise because conditional effects or covariate distributions in the full and selected samples differ in important ways. Other examples of collider bias include responder bias and biased analysis based on improper subgroups. Choi et al. provide numerous examples of selection bias in rheumatic disease research.
An interesting example in rheumatic diseases is discussed by Nguyen et al., who considered the smoking paradox wherein smoking is a risk factor for psoriatic arthritis, but has been reported to have a protective effect for psoriatic arthritis among psoriasis (Ps) patients; they explain this finding, conceived as arising from a collider bias, through mediation analysis. Specifically, by fitting Cox regression models conditional on confounders such as sex, age, BMI, alcohol intake and a history of trauma, their mediation analysis suggests that smoking is associated with an increased risk of psoriasis and an increased risk of psoriatic arthritis in the population (hazard ratio (HR) = 1.27; 95% CI: 1.19, 1.36). However, the direct effect on the risk of psoriatic arthritis is estimated to be \( HR = 0.96 \) (95% CI: 0.93, 1.00); thus, the primary reason for the smoking-PsA effect is an indirect causal effect through smoking increasing the risk of developing psoriasis, and psoriasis in turn increasing the risk of psoriatic arthritis. We describe the situation in more detail below.

Baker et al. investigate the finding that greater body mass appears to predict less radiographic progression of joint damage over 1-2 years, from the perspective of collider bias. Recently, selection bias has arisen in discussion of COVID-19 studies of hospitalized patients, where associations between the presence of underlying rheumatic diseases and COVID-related death were examined; here there is an important distinction between associations and causal effects in the broader population of individuals with COVID-19 and those admitted to hospital. The latter sub-population will have different distributions of confounders and effect modifiers so findings from analyses of hospitalized patients will not be externally valid for the full population.

Selection mechanisms can similarly have a substantial influence on findings in registry-based studies. Consider a binary \( W \) which indicates recruitment to a disease registry (i.e. \( W = 1 \) for an individual recruited to the registry and \( W = 0 \) otherwise). If selection to the registry depends on both \( X \) and \( Y \) then even if \( X \perp Y \) in the population, \( X \perp Y | W = 1 \) in the sample. Analysis of registry data necessarily involves conditioning on \( W = 1 \), so will suggest that \( X \perp Y | W = 1 \). The subtle point...
is that conditioning on \( W \) here is implicit simply through use of available data. This selection bias might be called a collider bias if viewing the process through Figure 2(b). If there exists an auxiliary variable \( U \) associated with recruitment and such that \( W \perp Y \mid U \) (see Figure 2(c)) then an analysis based on \( P(Y \mid X, U) \) should suggest \( Y \perp X \mid U \) since the variable \( U \) addresses the collider (selection) bias in this example. Choi et al. 28 discuss a situation where a variable \( U \) that is associated with the presence of osteoarthritis (OA) can alter the effect of a factor such as obesity on disease progression. The collider variable in this case is the presence of OA. It is consequently important in observational studies to collect information on factors associated with the selection process, rather than simply describing the composition of the available sample.

Further Remarks on the Smoking Paradox

As an illustration we consider the psoriasis-psoriatic arthritis smoking paradox discussed by Nguyen et al. 32. A binary indicator of smoking status (equaling 1 for a smoker and 0 for a nonsmoker) at a given time is denoted by \( X \) in Figure 3; for simplicity we exclude previous (but not current) smokers. The binary outcome variable \( Y \) indicates whether a person has PsA (\( Y = 1 \)) or not (\( Y = 0 \)) at that time, and the binary variable \( W \) indicates whether a person has psoriasis (\( W = 1 \)) or not (\( W = 0 \)). Covariates, which may include other risk factors for PsA or Ps, are represented by \( V \); although the graph does not show this, \( V \) may be associated with \( X \).

For simplicity we consider linear models

\[
E(Y \mid X, V, W) = P(Y = 1 \mid X, V, W) = \beta_0 + \beta_1 X + \beta_2 V + \beta_3 W, \tag{4.1}
\]

And

\[
E(W \mid X, V) = P(W = 1 \mid X, V) = \gamma_0 + \gamma_1 X + 2V. \tag{4.2}
\]
Conventional binary regression models would involve link functions to ensure that the probability lies in the interval [0,1] but we consider this simple model for convenience. We note that (4.1) implies that among persons with psoriasis,

\[
E(Y|X,V,W = 1) = P(Y = 1|X,V,W) = \beta_0 + \beta_1 X + \beta_2 V + \beta_3,
\]

and (4.1) and (4.2) together imply that in the general population,

\[
E(Y|X,V) = (\beta_0 + \beta_3 \gamma_0) + (\beta_1 + \beta_3 \gamma_1)X + (\beta_2 + \beta_3 \gamma_2)V.
\]

Thus the effect of smoking in persons with psoriasis (conditional on \(V\)) is \(\beta_1\), and in the general population (conditional on \(V\)) it is \(\beta_1 + \beta_3 \gamma_1\). The smoking paradox is that \(\beta_1 + \beta_3 \gamma_1 > 0\) whereas \(\beta_1 < 0\); see Table 3 of Nguyen et al.\(^{32}\) for example. Our \(\beta_1\) corresponds to their “direct” effect and \(\beta_3 \gamma_1\) to their “indirect” effect and they explain the paradox via mediation analysis (\(W\) is the mediating variable between \(X\) and \(Y\)) and the “collider” effect from conditioning on \(W = 1\) when considering persons with psoriasis.

We note that there are other factors which may bias estimation of these effects, so that any conclusions here are tentative. These include the fact that time is not thoroughly accounted for; all of \(Y\), \(X\) and \(W\) are time-dependent and simply considering a fixed time for each person is problematic. In particular, it is also necessary to condition on the disease and smoking history of each individual at the time their follow-up in the study begins. There may also be \(X - V\) interaction effects and a lag in the effect of smoking – these have not been considered. In short, the interplay between the time-varying factors \(Y(t), X(t), W(t)\) and \(V(t)\) here is sure to be complex and an analysis that accounts for all potential sources of bias is challenging.

Numerical Illustration of Selection Effects as a Collider Bias
To illustrate the selection effect here we consider the case where an exposure variable $X$ of interest is not associated with a response $Y$ ($X \perp Y$) and suppose for simplicity that both are continuous variables with $X \sim \mathcal{N}(0, 1)$ and $Y \sim \mathcal{N}(0, 1)$ in the study population. Suppose that certain individuals from the population are recruited to join a registry, and let $W = 1$ indicate that an individual is recruited, with $W = 0$ otherwise. Now suppose the probability a person is recruited depends on their $X$ and $Y$ values, with

$$
\log \left( \frac{P(W = 1 \mid X, Y)}{1 - P(W = 1 \mid X, Y)} \right) = \rho_0 + \rho_X X + \rho_Y Y.
$$

If $\rho_X = \rho_Y = 0$ then the registry involves simple random sampling from the population, but recruitment is typically outcome- and possibly exposure-related. Suppose for example, that $\exp(\rho_Y) = 5$ so an individual with a one unit higher value of $Y$ than another (with the same value of $X$) has a 5-fold higher odds of being recruited to the registry. Likewise let $\exp(\rho_X) = 4$ so an individual with a one unit higher values of $X$ than another with the same value of $Y$ has a 4-fold greater odds of being recruited. We set $\rho_0$ so that the probability of recruitment to the registry is 0.25. Figure 4(a) shows a scatter-plot of $Y$ versus $X$ for a population of 1000 individuals with the dots coloured blue for individuals recruited to the registry and red otherwise; the circular cloud of points reflects the independence of $X$ and $Y$ in the full population. The scatter plot of $Y$ versus $X$ based on the individuals in the registry in Figure 4(b), on the other hand, spuriously suggests a negative association between $X$ and $Y$. Fitting a standard linear regression model for $E(Y \mid X) = \beta_0 + \beta_1 X$ to the data in the registry (better represented by $E(Y \mid X, W = 1) = \beta_0 + \beta_1 X$) gives an estimated regression coefficient $\hat{\beta}_1 = -0.23$ with a standard error $\text{s.e.}(\hat{\beta}_1) = 0.02$ and a p-value $< 0.001$. This is an extreme scenario but it illustrates the importance of understanding and characterizing the ways in which individuals are selected in cross-sectional studies.

Selection effects also arise in randomized clinical trials when drop-out, non-compliance or treatment switching occur and attention is restricted to a sub-sample of randomized individuals.
Figure 1(b) depicts a clinical trial where $X \perp W$ due to randomization; both $X$ and $W$ exert an effect on $Y$, but one may aim to estimate a causal effect with or without conditioning on $W$. With drop-out let $C = 1$ indicate that an individual completes the study and $C = 0$ otherwise. A complete case analysis restricts attention to individuals who completed the study and implicitly conditions on $C = 1$. Thus while $W \perp X$ in Figure 1(b) due to randomization, $X \parallel W \mid C = 1$ in Figure 5. Conditioning on the collider ($C = 1$ indicating completion of the study) renders $W$ an active confounder despite randomization, and an analysis based on $P(Y \mid X, C = 1)$ will obscure the true relation between $X$ and $Y$. A proper analysis requires us to represent the persons who did not complete the study. This can be achieved by modeling drop-out as a function of observed auxiliary variables $U$ and if dropout is independent of $Y$ given those variables (so $C \perp Y \mid (W, X, U)$), this model can be used to reweight the observations to adjust for drop-out\(^\text{35}\). Another approach is to use multiple imputation\(^\text{36}\) involving an outcome model including such covariates.

Finally, we comment on the phenomenon of mediators, introduced in the previous smoking paradox illustration. Mediators are not sources of bias but their study is essential to understanding causal mechanisms. Figure 2(d) represents a setting in which the causal effect of $X$ on $Y$ is (partially) mediated through $W$. This means that a change in $X$ causes a change in the mean of $Y$ directly as represented by the $X \rightarrow Y$ relationship, and indirectly by the $X \rightarrow W$ and $W \rightarrow Y$ arrows. In such settings the overall (or total) effect of $X$ on $Y$ is an aggregate of the direct and indirect effects; see VanderWeele\(^\text{37}\). These are easily examined when linear models are applicable, but more difficult in other cases. In the smoking paradox illustration we saw how a linear relationship between $Y, X$ and $W$ and another linear relationship between $X$ and $W$ suggested that the effect of smoking on PsA was mediated by the presence of psoriasis; this produced a “total” effect of $X$ on $Y$ that was a sum of direct and indirect effects. As another example, in studying treatment effects in rheumatoid arthritis one may aim to study the extent to which biological therapy reduces the rate of joint damage over two years through a reduction of c-reactive protein. Recently de Vlam et al.\(^\text{38}\) studied the mediating
role of inflammatory markers (CRP, swollen joint count, Leeds enthesitis index, PASI score and an itch severity score) on the causal effect of Tofacitinib on pain reduction in psoriatic arthritis.

5. The Central Role of Time in Disease Processes

The dynamic nature of disease processes and associated factors impacts study selection, treatment initiation, and observation processes (e.g., clinic visits and loss to follow-up). It is therefore essential to generalize the concepts, notation and models discussed thus far to take time into account. In this case we use $Y(t)$, $W(t)$ and $X(t)$ to represent outcome, covariates and exposure (or treatment) variables at time $t$ since disease onset or some other time origin. In a given setting $X(t)$ might be fixed over time, as might some or all of the covariates in $W(t)$.

Two broad approaches to analysis of time-varying disease processes are (i) to use dynamic models that describe how the variables $Y(t)$, $X(t)$, $W(t)$ evolve over time, and (ii) to use models that focus on $Y(t)$ at a specific time $t$, conditional on $Y(0)$, $X(0)$, $W(0)$. Dynamic models are essential for a full understanding of disease processes and for prediction of future events conditional on a person’s current and past disease states. Effects in such models are by nature conditional, since they condition on prior history of disease and other factors. A fundamental example is the Cox model for failure time outcomes; here relative risks for exposures or treatments are expressed in terms of the hazard function, which represents the instantaneous probability of failure at time $t$, conditional on not having failed earlier. However, when we want to assess the causal effect of an intervention in a randomized trial, such models are problematic because by their very nature they condition on post-randomization events (e.g. being alive at time $t$) that may be related to treatment and other factors; this introduces time-dependent confounding since even if $X$ and $W$ are independent at the time of randomization ($t = 0$) they are no longer independent for persons alive at time $t$.

Issues discussed in earlier sections for “static” or cross-sectional settings are thus more complicated when time is involved. Conditional effects of treatment $X$ given covariates $W(t)$ and
previous disease history can be specified through process intensities, and are especially important for understanding process dynamics. Aalen et al. provide an excellent discussion of forms of causality, including what they term dynamic causality. In randomized trials there is, however, a desire to base a marginal treatment effect on a marginal process feature. Bühler et al. provide a thorough discussion of such effects. There is a large literature on methods for “causal” analysis of treatment effects in observational studies. As discussed in Section 2, they require strong unverifiable assumptions, including the absence of unobserved confounders and a positive probability of receiving each treatment for every individual. Confounding is often related to time and to dynamic markers of disease activity that are associated with symptoms which influence treatment decisions (and even engagement with the healthcare system) and are also associated with disease progression. In the case of PsA, for example, treatment assignment may be related to a biomarker such as ESR, which is also related to disease outcomes and responsive to treatment. Collider effects may arise in observational studies involving time in various ways, but the most important are through study recruitment and selection, and through dropout processes.

In controlled experimental studies, protocols specify the process by which individuals are selected, seen, and measurements are taken. As discussed in Section 2.4, complications can occur when individuals withdraw from a study. Although methods for dealing with this are available, there are challenges when the time of dropout depends on disease- or treatment-related events; failure to recognize this can lead to biased inferences concerning treatment effects. Biases may also arise in studies where participants are seen at intermittent visits, if the timing and frequency of visits is related to a person’s disease process. For recent discussions of such factors see Cook and Lawless and references therein.

Transportability of results from an observational study is more likely for detailed models and conditional effects that relate disease outcomes to prior disease history plus important covariates and treatment information. The assessment of interventions in observational studies is usually
challenging, however; in dynamic settings treatment assignment can be difficult to handle unless data are available for modeling the treatment decision; collecting such data should be a priority. We discuss this issue in a subsequent article.9

6. Enhancing Transportability and Replicability of Research

In the preceding sections we discussed various statistical challenges for disease history studies. Here we return to the issues of replicability and transportability, and ways they may be facilitated.

6.1 Replication and Study Designs

A primary consideration in any research study is a clear specification of the target population, study population and participant selection plan. The target population is the population of persons about whom inferences are desired, while the study population is a subset of the target population from which participants are drawn. In a study evaluating an experimental treatment the study population will be a subset of the population of individuals who may eventually be candidates for treatment. Such studies commonly exclude individuals for whom there may be safety concerns (e.g., pregnant women) or individuals with serious comorbidities that may jeopardize the chance that they will furnish useful information on the outcomes of interest. The distribution of key auxiliary variables will typically differ in study and target populations, so it is not surprising that average treatment effects may fail to replicate precisely if a new study is based on a different study population, or even the original target population. Conditional treatment effects may be more easily replicated, since they adjust for auxiliary variables.

Observational studies are inherently more varied and complex than most randomized controlled trials. There are various types of observational studies, which differ in the ways participants are recruited or selected. This includes i) selection according to a formal survey design involving probability sampling from a specified population;42 ii) random selection of persons in a
population satisfying certain conditions;\textsuperscript{43} iii) inception cohorts or disease registries that enroll persons at the time they have a specific event;\textsuperscript{44} iv) prevalent cohorts that recruit persons with a particular condition or in a particular life state; and v) registries or cohorts consisting of patients receiving care at a clinic or set of facilities.\textsuperscript{45} For i) - iv) the target and study populations may be fairly clear, but this can be much less so for v). Some studies may involve only the collection of retrospective data on participants, whereas others involve prospective data collection.

A study selection plan should describe how individuals are identified for possible recruitment to the study and list inclusion and exclusion criteria. Nevertheless, it may be hard for a new study to involve a similar set of participants. For example, a psoriasis study may screen a population via questionnaire, while another may seek referrals from family physicians, and another may recruit participants from specialty dermatology clinics. Although similar entry criteria may be specified for each study it is expected that the clinics would involve patients with a longer disease history and more severe manifestations.

Prospective studies of disease processes typically involve the collection of data over a prolonged period, and a study protocol should also be clear on how participants’ data is collected. This involves the specification of which clinical events, biomarkers, and auxiliary covariates are to be recorded, as well as the measurement processes and the frequency and timing of data collection. The logistics of data collection may vary substantially across studies because of differing environments and levels of resources, and so substantial variation across studies is common in spite of guidelines such as the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) criteria.

Thus, situations where it is feasible to closely replicate a previous study’s design are rather rare. Indeed, for cohort studies in which patients are recruited from a special population (e.g., persons receiving care at a specific hospital system or series of local facilities) or by a variety of different methods, this is usually impossible. For large cohort studies based on random samples from a target
population (e.g., Raina et al.\textsuperscript{42}) this may be possible to a degree, but such studies can be costly and duplicating them may not be considered a good use of resources.

Well-designed studies, combined with proper analysis that recognizes features of the study design and thus avoids biased inference, provide results that may be termed internally valid, in the sense that the inferences are transportable to the study population. Replicability refers more specifically to comparison of results, especially concerning effects: are they similar in two studies? It is important to recognize that some results may replicate, but others may not. As we have discussed, inferences about average effects are less likely to be similar across studies or populations than conditional effects. This should not be a cause for undue concern; a comparison of studies should include comparison of the distributions of auxiliary variables in the studies. Medical science progresses through studies that enhance understanding of disease processes and their treatment, which are by their complex nature resistant to simple (marginal) summarization. Good studies should seek to provide transportable conditional models and effects which have some degree of causal plausibility. We now discuss some aspects of analysis that, in combination with good design, facilitate this.

\section*{6.2. Some Aspects of Analysis}

A proper analysis of study data must take into account conditions imposed, by design or through random events, on the selection and follow-up of participants. Factors that may compromise internal validity of a study or complicate future assessments of replicability or transportability of results should be identified where possible; we give three examples that involve different aspects of a study.

First, in the selection of study participants, refusals are common among persons contacted. Collection of information on key attributes for all persons contacted facilitates assessment of the
representativeness of a study sample, and can shed light on comparisons with other studies. A second example is in clinical registries where persons have scheduled regular assessments, but may miss or significantly delay clinic visits. If the propensity to hasten or delay clinic visits is related to disease activity, failure to address this can bias estimates of disease progression or complication rates. In this situation it is important to obtain information about the reasons why a person visits the clinic and has data recorded, as well as disease-related information in the period since the most recent visit. A third example concerns adjustment for baseline factors in the analysis of observational studies where participants may have different disease durations and levels of recent disease activity upon entry. It is important to collect accurate information about this, as well as information that sheds light on why a person joined the study. An interesting illustration is in connection with the Women’s Health Initiative (WHI) randomized trial on the effects of hormone therapy (HT) on post-menopausal women. The WHI trial indicated an elevated risk of certain cardiovascular (CV) events for users of HT whereas earlier observational studies had indicated a lowered risk. The discrepancy was largely due to the fact that effects of HT were time-dependent and that the women in the WHI were followed from initiation of HT whereas many in the observational studies were enrolled after HT had been initiated. In addition, observational studies sometimes included women initiating HT before menopause. See Prentice et al. for a discussion of study designs and analyses.

We note that for all types of studies, descriptive statistics are often provided in analyses and articles to give some sense of the study population. However, this is typically done one variable at a time. More complete information on the joint distribution of key variables will facilitate assessments of transportability; this can be made available by sharing data from studies, as proposed in the movement for reproducible research.
6.3 Confounders and How they are Dealt With

We have stressed the importance of understanding the relationships between variables representing outcomes, interventions or exposures, and other covariates. Directed acyclic graphs (DAGs) offer a useful way of enumerating core variables, communicating beliefs about how they affect the process of interest, and discussing how they can be dealt with in analyses. Much of the discussion concerning DAGs and “causal” inferences does not adequately address the important role of time. Time is always present implicitly, even in static DAGs such as those in Figures 1 and 2, since an arrow from $X$ to $Y$, say, implies $Y$ is realised after $X$. However, when variables $Y(t)$, $X(t)$, $W(t)$ are time-varying it is also important to consider how current and past values of variables affect future values for them and other variables. The importance of measuring and considering confounders when assessing the effects of exposures or interventions is widely recognized but for a given study it can be challenging to provide a comprehensive list of potential confounders. When processes are dynamic and information is intermittently recorded this is more complicated; for example, unobserved events that occur between visits may affect clinical events and treatment decisions.

Although the better transportability and generalizability of conditional effects make them more appealing than average effects, there remains much interest in average effects, which superficially appear to be easily interpreted. As we discussed earlier, two ways of estimating average effects for $X$ in observational settings are to fit models for $Y$ given $X$ and $W$ and then to average over $W$ values, or to utilize weighting based on propensity scores. However, for dynamic processes where treatments may be assigned or terminated at various times, $W(t)$ frequently involves time-varying components such as biomarkers. It is not obvious how to average over such variables and many authors have discussed methods that seek to emulate random assignment of treatment over specified time periods – see Dickerman.47 We discuss such issues in Cook and Lawless.9
6.4 Comments on Real World Evidence

The availability of large databases on health care delivery and health outcomes has sparked interest in the use of so-called real-world data to obtain evidence of treatment effects (e.g., Crown\textsuperscript{48}). The FDA defines real world data (RWD) as "data relating to patient health status and/or the delivery of health care routinely collected from a variety of sources." We take "real world" to imply a broadly defined target population where interventions and treatment decisions are made according to some standard of care. This setting differs greatly from that of a clinical trial due to variation in covariate distributions, the likely presence of confounding variables, and the potential for selection or collider bias due to inclusion conditions for the database. For example, inclusion in the database and information on the health status of individuals therein may depend on the health status itself - individuals seek medical care when they are experiencing disease activity or progression. Effect heterogeneity in target populations implicit in real-world data may also be an issue, especially if there are interactions with variables not in the database. There are numerous other caveats concerning real-world data (RWD). For example, it is difficult to measure outcomes and other factors as accurately as in clinical trials and well-run observational studies; measurement error is common.

Much discussion about RWD has focussed on the estimation of average causal effects\textsuperscript{14,48,49} in spite of difficulties that we have discussed here. These issues suggest there are many problems in assessing or interpreting average effects with real-world data. A related concept is real-world evidence (RWE), which has been defined as "clinical evidence about the usage and potential benefits or risks of a medical product derived from analysis of RWD"\textsuperscript{50}. This does not mention either average or conditional causal effects explicitly, but once again there is more scope for assessing conditional effects, which as we have discussed, are more likely to be transportable.

7. Discussion
Transportability and replication are integral to scientific research, and also to the frequentist interpretation of estimates and hypothesis tests, which calibrates them by reference to conceptual repetitions of a study. Much of the discussion regarding “replicability crises” has concerned treatment effects and issues related to significance testing, selective inference, and publication bias. In observational settings careful consideration of the target population, the representativeness of the study sample, confounding, and the formulation of statistical models and analysis strategies that address these issues are central, and we have focused on them. We note that articles in scientific journals typically do not provide sufficient detail for an assessment of transportability; there have been calls for more rigorous statistical assessment as part of the review process for scientific journals (e.g., McNutt) and calls for more intensive efforts at statistical education.

In a seminal contribution to epidemiology and medical research, Hill offered guidance on when one might be justified to infer that an association seen in an observational study might reflect a causal effect. He does not use the term replication, but states that, among several other principles, “consistency” of findings across several studies involving “different persons” (read “populations”), “places, circumstances and times” would strengthen causal inferences. We hope our discussion has shed some light on factors influencing replicability, or at least comparability, of findings across different studies.

Analyses which condition on variables besides the one of interest can provide protection against confounding. The identification of confounding variables is a scientific exercise which requires understanding of the context in which exposures and outcomes arise. Elements of $W$ needn’t be statistically significant in outcome regression models for them to play a useful role in mitigating potential confounding effects. The inflammatory marker ESR is a good example of a potential confounder that may influence a clinical decision to prescribe treatments, and high values are known to be associated with progression of joint damage. Other potential confounders include...
We conclude with some take home messages.

- There is no substitute for good study design that provides cohorts representative of the study population, and that collects accurate data on clinical outcomes and on covariates that may be associated with outcomes. It is also important to obtain information on variables or factors related to a person’s inclusion in a study and, during the study, on factors related to the timing and accuracy of prospective data collection.

- A clear description of how a study sample or cohort was obtained is needed in order to facilitate assessment of internal or external validity (transportability to a study or to a target population).

- Careful analysis of observational data is needed when studying exposure or intervention effects. This includes specification of relevant covariates and an assessment of potential confounding effects. This may also raise awareness of covariates that are not available, in turn highlighting limitations of the study.
• The effect of observed confounding variables on the estimation of marginal effects should be mitigated by the use of methods such as regression adjustment, stratification, or the use of propensity scores. However, caution is warranted when reporting average causal effects in the presence of heterogeneity. Lunt et al.\textsuperscript{56} point out that different strategies for balancing covariates across exposure groups can lead to important differences in estimates of average causal effects; see also Stürmer et al.\textsuperscript{57}

• Conditional causal treatment or exposure effects expressed (i.e. conditional on covariates) are more transportable than average causal effects.

• Collider bias can arise from conditioning on a variable that alters the association between an exposure and a response. This often occurs when study participants are a selective subset of the study population. Addressing this involves modeling and analyzing the collider variable.

Much of our discussion was on situations where issues could be discussed using DAGs in a static environment. Even in this setting there is an implicit temporal aspect since for the causal interpretation of DAGs, a cause must precede an effect. For more complete modeling and analysis of dynamic disease processes, one must address factors such as time-dependent confounding, the occurrence and timing of observations in cohorts and registries, and disease-related loss to follow-up. In a companion article\textsuperscript{9} we elaborate on the importance of framing scientific questions and conducting analyses in ways that respect the selection and follow-up processes for a study, along with temporal aspects of disease processes and interventions.
References


Figure 1 Title. Some simple DAGs including an exposure $X$, covariates $W$, and an outcome $Y$.

Sub-title for Figure 1 a): A possible observational setting involving $X$, $W$ and $Y$.

Sub-title for Figure 1 b): A DAG for a setting where $X$ is randomized and $X \perp W$.

Figure 2 Title. Directed acyclic graphs depicting confounders, colliders, and mediators.

Sub-title for Figure 2 a): $W$ is a confounder.

Sub-title for Figure 2 b): $W$ is a collider.

Sub-title for Figure 2 c): Mitigating collider bias via an auxiliary variable.

Sub-title for Figure 2 d): $W$ is a mediator.

Figure 3 Title. A causal graph for discussion of the smoking paradox involving covariates ($V$), smoking status ($X$), Ps status ($W$) and PsA status ($Y$).

Figure 4 Title. An illustration of the collider bias in a hypothetical setting with continuous $X$ and $Y$ in a population and a registry.

Sub-title for Figure 4 a): Scatter plot of the response $Y$ against the exposure $X$ in the population of 10000 individuals; red dots represent unselected individuals and blue dots denote those recruited to the registry.

Sub-title for Figure 4 b): The scatter plot showing the relation between $Y$ and $X$ in those recruited to the registry depicting spurious evidence of a negative association.
Figure 5 Title. DAG for a randomized clinical trial (RCT) with drop-out where $C = 1$ signifies completion of the trial and $C = 0$ otherwise.

(a) A possible observational setting involving $X$, $W$, and $Y$.

(b) A DAG for a setting where $X$ is randomized and $X \parallel W$.

Fig. 1: Some simple DAGs including an exposure $X$, covariates $W$, and an outcome $Y$.

(a) $W$ is a confounder.

(b) $W$ is a collider.

(c) Mitigating collider bias via an auxiliary variable.

(d) $W$ is a mediator.

Fig. 2: Directed acyclic graphs depicting confounders, colliders, and mediators.

Fig. 3: A causal graph for discussion of the smoking paradox involving covariates ($V$), smoking
status \((X)\), Ps status \((W)\) and PsA status \((Y)\).

Fig. 4: An illustration of the collider bias in a hypothetical setting with continuous \(X\) and \(Y\) in a population and a registry

\[
\begin{array}{c}
W \\
\downarrow C \\
Y \\
\downarrow X
\end{array}
\]

Fig. 5: DAG for a randomized clinical trial (RCT) with drop-out where \(C = 1\) signifies completion of the trial and \(C = 0\) otherwise.