Recoverability of causal effects under presence of missing data: a longitudinal case study

Anastasiia Holovchak [1,*], Helen McIlleron [2], Paolo Denti [2],
Michael Schomaker [3,4]

[1]Seminar für Statistik, ETH Zürich, Rämistrasse 101, 8092 Zürich, Switzerland
[2]Division of Clinical Pharmacology, Department of Medicine, Faculty of Health Sciences, University of Cape Town, 7935 Observatory, Cape Town, South Africa
[3]Department of Statistics, Ludwig-Maximilians Universität München, Ludwigstraße 33, 80539 München, Germany
[4]Centre for Infectious Disease Epidemiology and Research, University of Cape Town, Anzio Road, 7925 Observatory, Cape Town, South Africa

*Corresponding author: Seminar für Statistik, ETH Zürich, Rämistrasse 101, 8092 Zürich, Switzerland
Email: anastasiia.holovchak@stat.math.ethz.ch

SUMMARY

Missing data in multiple variables is a common issue. We investigate the applicability of the framework of graphical models for handling missing data to a complex longitudinal pharmacological study of children with HIV treated with an efavirenz-based regimen as part of the CHAPAS-3 trial. Specifically, we examine whether the causal effects of interest, defined through static interventions on multiple continuous variables, can be recovered (estimated consistently) from the available data only. So far, no general algorithms are available to decide on recoverability, and decisions have to be made on a case-by-case basis. We emphasize the sensitivity of recoverability to even the smallest changes in the graph structure, and present recoverability results for three plausible missingness-directed acyclic graphs (m-DAGs) in the CHAPAS-3 study, informed by clinical knowledge. Furthermore, we propose the concept of a "closed missingness mechanism": if missing data are generated based on this mechanism, an available case analysis is admissible for consistent estimation of any statistical or causal estimand, even if data are missing not at random. Both simulations and theoretical considerations demonstrate how, in the assumed MNAR setting of our study, a complete or available case analysis can be superior to multiple imputation, and estimation results vary depending on the assumed missingness DAG. Our analyses demonstrate an innovative application of missingness DAGs to complex longitudinal real-world data, while highlighting the sensitivity of the results with respect to the assumed causal model.

KEYWORDS: causal effect; longitudinal study; missing data; missingness DAG; multiple imputation.

1. INTRODUCTION

Missing data are a common issue in biomedical research. In particular, longitudinal epidemiological studies, where data are collected consecutively at several points in time, tend to suffer from missing data on multiple variables. One common assumption which is often made regarding the missing data mechanism is the so-called *missing at random* (MAR) assumption. If this assumption holds,

statistical methods for consistent estimation are available, for example, multiple imputation or inverse probability weighting (Rubin 1976; Mohan and Pearl 2014).

However, if multiple variables are missing simultaneously, it is often difficult to justify whether the MAR assumption holds. This is both because the classic MAR definition is event- rather than variable-based, and arguing for or against conditional independencies of *events* is practically challenging, especially in longitudinal settings (Mohan and Pearl 2021). Moreover, even though MAR is the weakest known condition under which the missingness process can be ignored (i.e. dealt with using the observed data), it is only a sufficient, but not a necessary condition for unbiased estimation; this means that under a *missing not at random* (MNAR) scenario, it is unclear if and how a target estimand can be recovered (estimated consistently).

To tackle these and other challenges, Mohan et al. (2013) and Mohan and Pearl (2021) proposed an alternative causal graph-based framework, in which knowledge and assumptions about *reasons* for missingness are encoded in relationships between *variables* and missing data indicators. This framework is very general and can be used to evaluate whether estimands can be recovered given a correct causal missingness model (missingness-directed acyclic graph, m-DAG).

However, currently there is no general algorithm that can be used to establish recoverability for arbitrary settings. Therefore, identification and estimation strategies have to be developed on a case-by-case basis for each particular scenario. This poses the question of how useful causal m-DAGs are practically.

Moreno-Betancur et al. (2018) have convincingly argued that one should thus develop canonical m-DAGs for recurring settings, such as for point-exposure study designs in epidemiology where missing data in outcome, exposure, and confounders are caused by some "standard" mechanisms. They demonstrated the usefulness of this approach in a study investigating the relationship between maternal mental illness and child behavior.

While causal missingness graphs, and their canonical versions, are a major advancement for causal inference research under missing data, their actual applicability has yet to be demonstrated in complex longitudinal studies. It is unclear how well knowledge on missing data can actually be collected, then integrated in a realistic causal graph, and how difficult the mathematical exercise of establishing identification and recoverability results in such a complex, yet realistic setting is. Can m-DAGs make their way from blackboards to actual applications?

Recent work, such as by Balzer et al. (2021) and Nugent et al. (2024), has advanced the field by using causal graphs and corresponding estimators to address recoverability and estimation of causal effects in longitudinal settings subject to complex missingness and dependence. These studies focus on reducing bias and improving efficiency by controlling for missing outcomes when estimating intervention effects. Their frameworks, both relying on a Two-Stage TMLE approach, are particularly useful for handling data in complex cluster randomized trial settings.

In this work, we focus on the explicit modeling of the missingness mechanism, not only for the outcome but for all relevant variables in the study. This involves developing a causal missingness model, explaining its motivation through a variable-based taxonomy, and then demonstrating whether identification is possible and what estimation strategies could be employed. We present an innovative application of missingness DAGs to a longitudinal study. Specifically, we investigate whether the derivation of identification results in a longitudinal setting is feasible, how volatile those results are, and how complex deriving those results can be.

In addition, we conduct simulation studies aligned with the data example to, firstly, verify the theoretical results on the recoverability of the desired causal query, and secondly, to quantify the extent of bias in settings of special interest. Our simulations and theoretical considerations demonstrate that causal diagrams help to explicitly guide the process of decision-making about whether a parameter of interest can be consistently estimated from the available data. In practice, MAR is often assumed and multiple imputation is performed, which in many cases may lead to biased estimation results. On the other hand, we demonstrate that even in complex longitudinal study settings, there are MNAR scenarios for which available case analysis leads to valid estimation results, whereas multiple imputation does not.

We sought to answer these questions by applying the framework proposed by Mohan and Pearl (2021) to answer a topical question in HIV-related pharmacoepidemiology, which is explained below in Section 2. We (i) investigate how clinical knowledge on reasons for our missing data can be best collected and integrated into a realistic causal missingness graph, (ii) derive recoverability results for our target estimand, (iii) discuss estimation strategies for it, (iv) evaluate the sensitivity of our results with respect to the assumed causal model, and (v) seek for structures that may be helpful in establishing results for future studies.

This paper is structured as follows. We introduce the motivating question in Section 2, followed by the theoretical framework in Section 3. In Section 4, analyze the data from the complex longitudinal pharmacological study CHAPAS-3, followed by the simulation study in Section 5. We conclude in Section 6.

2. MOTIVATING STUDY

Our motivating data analysis comes from the Children with HIV in Africa–Pharmacokinetics and Adherence/Acceptability of Simple antiretroviral regimens randomized trial (CHAPAS-3). The study enrolled 478 children with HIV, between 1 mo and 13 yrs of age, in 4 sites in Uganda and Zambia (Mulenga et al. 2016). Children enrolled into the trial received combined antiretroviral therapy, i.e. one non-nucleoside reverse transcriptase inhibitor (efavirenz or nevirapine) and two nucleoside reverse transcriptase inhibitors (abacavir, stavudine, or zidovudine—which were the randomized components– and lamivudine).

We are interested in determining target concentrations using data from 125 children treated with efavirenz (EFV). Efavirenz is used not only in children but also adults, though dosing recommendations (between 200 and 600 mg) depend on weight and age. Due to their different metabolic profiles and adherence patterns, patients with the same efavirenz dose may have different concentrations, conferring different protection against viral replication. Knowledge about concentrations is often used to derive dosing recommendations using population PK models (Bienczak et al. 2016). It is typically recommended that EFV concentrations between 1 and 4 mg/L should be achieved (at 12h after the dose was given, C12h). This is because lower concentrations may be insufficient to guarantee viral suppression and thus effective treatment, while higher concentrations may lead to toxicity negatively affecting the central nervous system (Schomaker et al. 2024). Our target estimand is thus the causal concentration-response curve (CCRC) at each follow-up visit, i.e. *we are interested in how the counterfactual probability of viral failure at time t varies as a function of possible prior concentration trajectories.*

Our analysis makes use of the data from Bienczak et al. (2016). We recently discussed statistical approaches on estimating the CCRC (Schomaker et al. 2023), but did not discuss aspects of missing data. In this manuscript, we extend the above study by developing a causal missingness graph (informed by the pediatrician's and trial team's knowledge) and derive identification and estimation approaches for our estimand of interest. More details on the analysis are given in Section 4.

3. FRAMEWORK

Missingness DAGs provide a natural extension of causal DAGs under the presence of missing data. Consider a DAG $\mathbf{G} = (\mathbf{V}, \mathbf{E})$ with a set of nodes $\mathbf{V}$, $|\mathbf{V}| = n$, which can be separated into two subsets $\mathbf{V}_o$ and $\mathbf{V}_m$, corresponding to sets of *completely observed* and *partially observed* variables. For each variable X_i, $i \in \{1, ..., n\}$, from the subset $\mathbf{V}_m$, a binary missingness indicator variable M_{X_i}

$$M_{X_i} = \begin{cases} 1 & \text{if } X_i \text{ is missing,} \\ 0 & \text{otherwise} \end{cases}$$

is introduced in the missingness DAG to depict causal relationships with other relevant variables. In the following, we refer to the set of missingness indicator variables as M. In addition, the setup allows for the existence of a latent (unmeasured) variable set $\mathbf{U}$.

In this work, we make use of the framework introduced by Mohan and Pearl (2021). We define $\mathbf{G}_c$ as the *complete-DAG (c-DAG)* (as in Moreno-Betancur et al. (2018)), and $\mathbf{G}$ as the *missingness DAG (m-DAG)*. The c-DAG $\mathbf{G}_c$ includes only the variable set $\mathbf{V} := \mathbf{V}_o \cup \mathbf{V}_m \cup \mathbf{U}$ that is relevant for identification considerations with respect to the (causal) parameter of interest (assuming no missingness in the variables in $\mathbf{V}_m$). The m-DAG $\mathbf{G}$, however, also includes the set of missingness indicators $\mathbf{M}$, and may additionally consist of a set of auxiliary variables, denoted as $\mathbf{Z}$, with variables in $\mathbf{Z}$ causing missingness in the variables in $\mathbf{V}_m$, but not being a cause of the variables from the set $\mathbf{V}$. Note that variables from both $\mathbf{V}$ and $\mathbf{Z}$ may be a cause of missingness in the variables from $\mathbf{V}_m$.

Mohan et al. (2013) and Mohan and Pearl (2021) show that causal diagrams can be used as a powerful tool for the identification and classification of missingness mechanisms. It is an important finding that the conventional taxonomy of missing data (Rubin 1976) can be translated (with some minor changes) into the context of causal diagrams, which is the focus of this work. However, it is essential to mention that the missingness taxonomy definitions proposed by Rubin (1976) and Mohan et al. (2013) are, in general, not equivalent. Note that Rubin's framework of missing data is defined on the *record-based level*, which makes it particularly difficult to apply in practice. Another issue that is often ignored is that the "realized" MAR definition as proposed by Rubin (1976) is not required to hold for all possible values of the missingness indicator M, but only for those present in the data. This makes clear that such a kind of definition does not hold for a data-generating distribution in general, but only for a single data sample, as different missing data patterns may arise for samples resulting from re-running of the experiment (Schafer and Graham 2002; Tian 2015). To overcome this issue, Seaman et al. (2013) distinguishes between two types of MAR—*realized* MAR and *everywhere* MAR. The latter is a "stronger" definition in the sense that conditional missingness distribution relates to the missing entries for all realized and unrealized patterns of missing data, and not only to the one specific data sample which has been observed.

Mohan et al. (2013) propose an alternative framework of missing data taxonomy, which is based on graphical models. The two main advantages of the graph-based method are an explicit encoding of dependencies on the variable level (and not record-based), and also depiction of causal mechanisms which are causing missingness (Tian 2015). Note that m-DAGs represent both the data-generating mechanism and the missing data mechanism, both in a causal manner.

Let $\mathbf{G}$ be an m-DAG over a set of variables $\mathbf{V}_o \cup \mathbf{V}_m \cup \mathbf{U} \cup \mathbf{M}$, where $\mathbf{V}_o$ and $\mathbf{V}_m$ denote the sets of completely and partially observed variables, correspondingly, $\mathbf{U}$ is a set of latent (unobserved) variables, and $\mathbf{M}$ is the set of missingness indicators. We denote the corresponding (joint) data distribution as $P(\mathbf{V}_o, \mathbf{V}_m, \mathbf{U}, \mathbf{M})$, and assume the distribution to be *faithful* (Pearl and Verma 1990) with respect to $\mathbf{G}$. Broadly speaking, faithfulness requires that the joint distribution $P(\mathbf{V}_o, \mathbf{V}_m, \mathbf{U}, \mathbf{M})$ satisfies all the conditional independence relationships encoded by the DAG, and only those relationships. In other words, one assumes that all observed conditional independencies follow from the graphical structure, and not from other reasons (such as deterministic relationships between variables).

The missingness mechanism is commonly characterized in terms of the conditional distribution of $\mathbf{M}$ given $(\mathbf{V}_o, \mathbf{V}_m, U)$, say $p(\mathbf{M}|\mathbf{V}_o, \mathbf{V}_m, U)$. It has to be assumed that any missingness indicator from the set $\mathbf{M}$ is not a parent of any variable from $\mathbf{V}_o \cup \mathbf{V}_m \cup U$. We emphasize once more that in the graph-based missing data framework, we work with variable-based and not the record-based definitions of the missingness mechanisms.

1. **Missing completely at random (MCAR).** If missingness occurs randomly and is assumed not to be caused by any variable in the model, missing or observed, we write

$$p(\mathbf{M}|\mathbf{V}_o, \mathbf{V}_m, \mathbf{U}) = p(\mathbf{M}),$$

which means that the conditional distribution of the missingness mechanisms given the variables in the data set and, possibly, also the latent variables, is equal to the marginal distribution of the missingness mechanism. This corresponds to the (unconditional) independence statement $(\mathbf{V}_o, \mathbf{V}_m, \mathbf{U}) \perp\!\!\!\perp \mathbf{M}$.
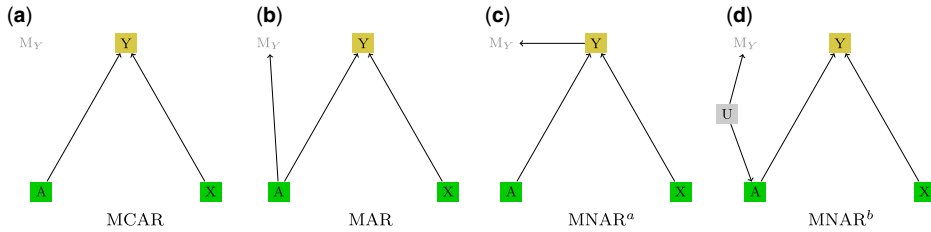
Figure 1: Example m-DAGs depicting three different types of missingness mechanisms: (a) MCAR, (b) MAR, (c) MNAR, (d) MNAR. Variables A and X are fully observed and variable Y is partially observed. We assume Y to be the disease indicator, A the type of treatment, and X the patient age. M_Y is the missingness indicator for the outcome Y.

In terms of an m-DAG, if the joint distribution $p(\mathbf{V}_o, \mathbf{V}_m, \mathbf{U}, \mathbf{M})$ is faithful with respect to the graph $\mathbf{G}$, this means that there are *no edges between the* $\mathbf{M}$ *variables and variables in* $\mathbf{V}_o$ *and* $\mathbf{V}_m$, and parents of missingness indicator variables from $\mathbf{M}$ can only be other variables from $\mathbf{M}$.

A commonly used example of the MCAR mechanism is an accidental technical error in an electronic medical record, which leads to a loss of data on the disease indicator, compare with Fig. 1a.

2. **Missing at random (MAR).** In this case, the missingness mechanism is assumed to depend on the fully observed variables only, but is not allowed to depend on the missing data values or latent variables. In terms of the conditional distribution of $\mathbf{M}$, this corresponds to

$$p(\mathbf{M}|\mathbf{V}_o, \mathbf{V}_m, \mathbf{U}) = p(\mathbf{M}|\mathbf{V}_o).$$

The conditional independence statement that holds in this case is $(\mathbf{V}_m, \mathbf{U}) \perp\!\!\!\perp \mathbf{M}|\mathbf{V}_o$, which in terms of m-DAGs (if the joint distribution is faithful to $\mathbf{G}$) means that *variables from* $\mathbf{M}$ *are not allowed to have any parents from the sets* $\mathbf{V}_m$ *or* $\mathbf{U}$, *but only from* $\mathbf{V}_o$ *and other* $\mathbf{M}$ *variables.*

For example, if a disease indicator is missing for some patients, and missingness for some reason depends on the completely observed treatment only, one says that the missing data underlies the MAR mechanism, as depicted in the example in Fig. 1b.

3. **Missing not at random (MNAR).** This category of missingness is most general. Data that cannot be classified as MCAR or MAR fall in the MNAR category. In this case, the conditional distribution of $\mathbf{M}$, $p(\mathbf{M}|\mathbf{V}_o, \mathbf{V}_m, \mathbf{U})$, cannot be simplified. In an m-DAG (again, assuming faithfulness), if at least one $\mathbf{M}$ variable has a parent which is a latent variable $\mathbf{U}$ or is any of the partially observed variables from the set $\mathbf{V}_m$, then the missing data mechanism is MNAR.

A common example of MNAR is when a variable is causing its own missingness, e.g. if the missingness of the disease status entry depends on the disease status itself, as this is the case for the m-DAG in Fig. 1c. For instance, the presence of a disease is likely to increase the chances of disease status being recorded due to the associated medical follow-ups and tests.

Another practically very relevant example is presented in Fig. 1d. It shows a case where missingness is driven by latent variables, which may potentially also be a cause for an observed variable. For example, socio-economic status (SES) may not be recorded in a medical study, but has an impact on both the missingness in the disease indicator due to reduced healthcare visits and the type of treatment an individual receives, since those with higher SES may have access to better healthcare facilities and more advanced treatment options.

Note that both missing data taxonomies are equivalent if (i) observations are independent, and (ii) missingness indicators are conditionally independent (Schomaker 2020). In this case,

the conditional distribution from the graph-based MAR definition can be written as a product of conditional distributions for single missingness indicators as well as for single observations, and this then leads to the equivalence of both definitions.

3.1. Recoverability of target quantities

In general, recoverability is a property of the target quantity/parameter θ (and the probability distribution P_X) which states whether this quantity can be estimated consistently from the available data. To decide on recoverability, one needs to know the independence statements that hold in the joint distribution. However, under the Markovness and faithfulness assumptions, all conditional independence statements present in the data can be directly read off from the m-DAG **G**, and recoverability can therefore be considered as a property of the pair $\{\theta, \mathbf{G}\}$ (Mohan and Pearl 2021). Note that the term *identifiability* is used when assessing whether a causal query can be expressed as a function of the observational distribution, whereas the *recoverability* concept is used to describe whether a parameter (not necessarily in a causal context) can be expressed as a function of the available data distribution (Moreno-Betancur et al. 2018).

Moreno-Betancur et al. (2018) describe three main conditions required for recoverability of a target parameter θ: consistency and well-defined interventions, positivity, and conditional independence conditions. In this context, consistency is a property of central relevance. It says that the factual treatment value coincides with the counterfactual outcome.

It has to be emphasized that many types of target parameters θ may be of interest, and different missingness patterns may occur, which makes it impossible to derive a general automatized algorithm allowing a decision about the recoverability of any specific θ. Therefore, the authors work on "canonical" scenarios which are most typical for epidemiological studies. Mohan et al. (2013) and Mohan and Pearl (2021) present some theoretical results on recoverability of joint and conditional distributions. In this work, we only briefly provide intuition on how recoverability works, focusing on aspects that are relevant for our own identification strategies. The main goal is to exploit conditional independencies between the variables of interest and the missingness indicators (which can be read off from the d-separation statements that hold in the graphs) in order to be able to condition on the missingness indicator variables (i.e. on $M_i = 0$ for some i). This way, observed variable values will be sufficient for consistent estimation of the target quantity.

In the following, we focus on the recoverability of causal effects. Indeed, necessary and sufficient conditions exist for recoverability of causal effects (Mohan and Pearl 2021). Under the presence of missing data, a *necessary condition for recoverability* of a causal effect is the identifiability of this effect from the c-DAG of interest. The causal effect is *identifiable* from a graph **G** if the interventional distribution can be determined uniquely from the observed-data distribution alone (Tian and Pearl 2002; Pearl 2009). Shpitser and Pearl (2006) provide a sound and complete algorithm for conditional causal effect identification (IDC), which, for any causal effect, can be used for determining identifiability and also for generation of an expression for the interventional distribution in the case of an identifiable effect. Tikka and Karvanen (2017) implemented the IDC algorithm in the R-package causaleffect. Using the IDC algorithm, we get an estimand of the causal query whenever identifiability holds for the causal effect. In turn, a sufficient condition for recoverability is that the (identified) estimand is recoverable from the missingness graph. This, as mentioned before, has to be decided on a case-by-case basis.

Next, an example of recoverability procedure of a causal effect for a simple longitudinal study with two measurement points is presented. The example is inspired by Mohan and Pearl (2021, Section 3.5). Consider the m-DAG in Fig. 2, which is a model of a simple two-time-point longitudinal study with attrition. The variables A_1 and A_2 correspond to sequential treatment, X_1 and X_2 are the side effects, and Y_1 and Y_2 model some health outcomes. The causal effect of interest is $P(Y_2; do(A_1, A_2))$, which is the impact of the two sequential treatments on the outcome at the second study time point. We can see that the partially observed variables X_1 and X_2 are causing their own missingness, which means that the missingness mechanism is of the MNAR type. However, even in this case, it can be shown that the causal effect of interest is recoverable using sequential
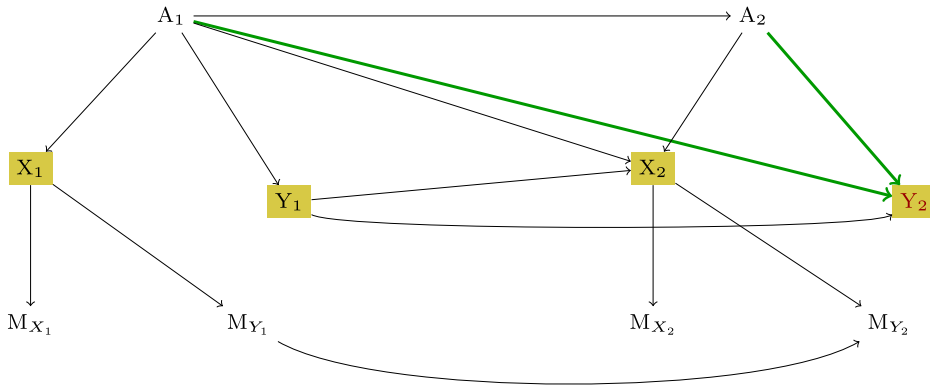
Figure 2: An m-DAG depicting the MNAR missingness mechanism in a simple longitudinal study. Health outcomes at two study time points are denoted as Y_1 and Y_2, sequential treatment variables as A_1 and A_2, and side effects as X_1 and X_2.

factorization (Mohan and Pearl 2021). First, with the help of the IDC algorithm, an expression for identifying the causal effect from the observable data is provided (for the case as if there had been no missingness, which is indicated in terms of potential outcomes):

$$P(Y_2^{M=0}; do(A_1, A_2)) = \sum_{Y_1} P(Y_2^{M=0}|A_1, A_2, Y_1^{M=0}) \times P(Y_1^{M=0}|A_1).$$

Note that the potential outcome notation $(\cdot)^{M=0}$ is explicitly used only for the variables observed incompletely. In the next step, it has to be decided on the recoverability of the two conditional distributions $P(Y_2^{M=0}|A_1, A_2, Y_1^{M=0})$ and $P(Y_1^{M=0}|A_1)$.

$$P(Y_2^{M=0}|A_1, A_2, Y_1^{M=0}) = P(Y_2^{M=0}|A_1, A_2, Y_1^{M=0}, M_{Y_1} = 0, M_{Y_2} = 0) \qquad (Y_2 \per\!\!\!\perp \{M_{Y_1}, M_{Y_2}\}|A_1, A_2, Y_1)$$

$$= P(Y_2|A_1, A_2, Y_1, M_{Y_1} = 0, M_{Y_2} = 0) \qquad \text{(by consistency)}$$

$$P(Y_1^{M=0}|A_1) = P(Y_1^{M=0}|A_1, M_{Y_1} = 0) \qquad (Y_1 \per\!\!\!\perp M_{Y_1}|A_1)$$

$$= P(Y_1|A_1, M_{Y_1} = 0) \qquad \text{(by consistency)}$$

Formally, it is true that $\{A_1, A_2, Y_1\}$ is the minimal set for which

$$Y_2 \per\!\!\!\perp \{X_1, X_2\}|A_1, A_2, Y_1.$$

Analogously, A_1 is the minimal set for which holds

$$Y_1 \per\!\!\!\perp X_1|A_1.$$

Further, $P(Y_2|A_1, A_2, Y_1)$ satisfies $Y_2 \per\!\!\!\perp \{M_{Y_1}, M_{Y_2}\}|A_1, A_2, Y_1$, whereas for $P(Y_1|A_1)$ it holds that true $Y_1 \per\!\!\!\perp M_{Y_1}|A_1$. Then, applying the sequential factorization technique, both factors can be recovered as $P(Y_2|A_1, A_2, Y_1, M_{Y_1} = 0, M_{Y_2} = 0)$ and $P(Y_1|A_1, M_{Y_1} = 0)$, correspondingly.

This shows that the causal effect of interest can be recovered from the available data only, despite the fact that the underlying missingness mechanism is MNAR.

After a causal effect of interest has been identified as a function of the observed data, (and after it has been further ensured that and how it can be recovered in the presence of missing data), the causal

effect can be estimated using the common techniques, e.g. (parametric) G-computation (Robins 1986), inverse probability weighting (Hernán and Robins 2006; Hernan and Robins 2020b), or targeted maximum likelihood estimation (TMLE) (van der Laan and Rose 2011).

3.2. Closed missingness mechanism

Next, an interesting special case of a missingness mechanism, which we refer to as *closed missingness mechanism*, is discussed. We present the recoverability results for joint and conditional distributions under the closed missingness mechanism.

Definition 1 (Closed missingness) Consider a c-DAG $\mathbf{G}_c$ with the node set $\mathbf{V} = \{X_1, X_2, ..., X_n\}$, $|\mathbf{V}| = n$. Without loss of generality, assume that $\mathbf{V}_o = \{X_1, ..., X_k\}$ and $\mathbf{V}_m = \{X_{k+1}, ..., X_n\}$ for some $k \leq n$. The corresponding m-DAG consists of the set $\mathbf{V}$, the missingness indicator variable set $\mathbf{M} = \{M_{X_{k+1}}, ..., M_{X_n}\}$, and possibly a variable set $\mathbf{Z}$ containing auxiliary variables that are causes of missingness. The missingness mechanism is called closed if there is no path between any $V_i \in \mathbf{V}$ and any $M_{V_i} \in \mathbf{M}$, $i \in \{k+1, ..., n\}$.

In other words, the missingness mechanism is closed if only the auxiliary variables from Z are causes of missingness. From a practical standpoint, this type of missingness mechanism is common in clinical, epidemiological, and pharmacological studies. The variables of interest in these studies are often of biological and medical nature (e.g. medication/therapy type, medication dose, blood values), whereas the causes of missingness in such variables are typically related to external factors such as technical issues with medical devices and the socioeconomic status of study participants.

Corollary 1 Consider an m-DAG $\mathbf{G}$ depicting a closed missing mechanism, and the general setting as in Definition 1. The joint distribution $P(\mathbf{V}_o, \mathbf{V}_m)$, and therefore also any marginal or conditional distribution, is recoverable from the available cases.

Proof. As there is no path between any $V_i \in \mathbf{V}$, $i \in \{1, ..., n\}$, and any $M_{V_i} \in \mathbf{M}$, $i \in \{k+1, ..., n\}$, V_i and M_{V_i} are d-separated for any i without conditioning on any other variables, and therefore $V_i \perp\!\!\!\perp M_{V_i}$ holds true. □

Note that the variables from $\mathbf{Z}$ may also be completely and incompletely observed, or even unobserved or latent, but this plays no role for the variables of interest from the set $\mathbf{V}$.

The concept of a "closed missingness mechanism" can lead to a useful practical rule of thumb, especially under MNAR. If a researcher determines that missingness is likely caused by unmeasured factors, they may initially conclude that the data are MNAR, making both imputation and available case analyses invalid. However, by sketching a c-DAG and m-DAG (incorporating missingness indicators), if no arrows are found from the c-DAG to the m-DAG, they can conclude that an available or complete case analysis is valid, despite MNAR. Intuitively, if the causes of missing values are unrelated to the variables necessary for identification, then relying solely on complete cases might be preferable to imputation.

This approach underscores a crucial practical insight: even in the presence of MNAR, if the missingness mechanism is closed, it does not impact the validity of complete case analysis. Thus, researchers can confidently use complete case data when the missingness indicators are isolated from the primary variables of interest.

4. DATA ANALYSIS
4.1. Study setting, measurements, and estimand

As described in Section 2, our data comes from the study of Bienczak et al. (2016) and includes 125 children with HIV who were enrolled in the CHAPAS-3 trial and treated with an efavirenz-based regimen. We consider the trial's follow-up visits at 6, 36, 48, 60, and 84 wks.

Our scientific goal is to estimate causal concentration response-curves. That is, we are interested in the counterfactual probability of a viral load (VL) > 100 copies/ml (which is considered to be a viral failure) at 36 and 84 wks if children had efavirenz concentrations (12h after dose) of x mg/L at each follow-up visit, where x ranges from 0 to 10 mg/L. Missing data in the outcome (VL), efavirenz exposure (EFV) and time-dependent confounders (weight, adherence) complicates this exercise.

Measured baseline variables include sex, age, the nucleoside reverse transcriptase inhibitors drug (NRTI) component of the treatment regimen and weight. Moreover, we include knowledge on the metabolism status ("genotype," slow, intermediate, extensive) related to the single nucleotide polymorphisms in the CYP2B6 gene, which is relevant for metabolizing evafirenz and directly affects its concentration in the body. In addition available data include knowledge on caregiver's beliefs in the necessity of medicine (BMQ, Abongomera et al. 2017), as well as socio-economic indicators (SES). Measured follow-up variables are weight, adherence (measured through memory caps, MEMS), and dose.

Clinical assessments were made at every visit, viral loads were measured at all time points except week 6, efavirenz levels were measured at all assessments other than week 48, and assessment of adherence through MEMS primarily at weeks 36, 48, and 60 as both funding constraints and practical considerations did not allow its consecutive implementation.

4.2. Development of the causal model

Figure 3 contains our proposed causal model. The causal graph contains (i) variables important for identifying the effect of interest (i.e. the impact of EFV on viral failure, c-DAG, bottom). We summarized (ii) clinician's knowledge on why data are possibly missing in a m-DAG using binary missingness indicators (top, pink shading) for relevant variables with missing data (EFV, elevated viral load (VL), adherence (MEMS), weight). The corresponding non-parametric structural equation models are given in Appendix A.

4.2.1. c-DAG

The c-DAG summarizes our knowledge and assumptions, described in more detail in Schomaker et al. (2023). Briefly, the primary cause of viral failure is subtherapeutic efavirenz concentration ($EFV_t \rightarrow VL_t$). The concentration itself depends on the administered dose ($Dose_t \rightarrow EFV_t$), adherence to treatment ($MEMS_t \rightarrow EFV_t$) and how fast the drug is cleared, determined—amongst other—by the 516G and 983T polymorphisms in the CYP2B6 gene ($Genotype \rightarrow EFV_t$). Both weight and MEMS are assumed to be time-dependent confounders. Co-morbidities (CoMo), which we defined to include tuberculosis, pneumonia and other AIDS-defining events are reflected in the DAG, although they are expected to be less frequent in our population, as children with active infections, those being treated for tuberculosis and with severe laboratory abnormalities at screening, were not enrolled into the study, and most children were in relatively good health (e.g. median CD4 cell percentage is 19%).

4.2.2. m-DAG

Main m-DAG ($\mathbf{G}_{main}$): The development of the missingness causal model is a novel contribution of our paper. Reasons for missing data were obtained from the trial team and the pediatricians. Those reasons are represented by arrows leading into the missingness indicators and include (i) technical issues with the memory caps (e.g. broken containers) or in obtaining or analyzing a blood sample (TI, unmeasured); (ii) missed visits (MV), which are likely related to socio-economic status of caregivers (SES, measured), beliefs and attitudes toward medicine (BMQ, measured) and other behavioral factors (BHV, unmeasured). As almost all children depend on their caregiver to arrive at their appointment, we assume that the children's age does not affect the probability of a missed visit.

Sensitivity of the main causal missingness model ($\mathbf{G}_{alt1}$): Although the clinicians did not report any other possible reasons for missingness, we added another speculative reason for missed visits to explore the implications of deviations from the assumed causal model: health status of the child, captured indirectly through elevated viral load. This potential reason is represented by the blue dashed arrows in the m-DAG ($\mathbf{G}_{alt1}$). In the discussion, we also mention another m-DAG ($\mathbf{G}_{alt2}$), where $\mathbf{G}_{main}$ is extended with additional arrows SES $\rightarrow$ MEMS$_t$, $t \in \{6, 36, 48, 60, 84\}$.

Following the introduced notations, and under the assumption of $\mathbf{G}_{main}$ or $\mathbf{G}_{alt1}$, the sets of completely and partially observed variables, auxiliary variables, and missingness indicators are defined as follows:

- $\mathbf{V}_o = \{Age, Genotype, Sex, NRTI, CoMo_t, Dose_t\}$
- $\mathbf{V}_m = \{MEMS_t, Weight_t, EFV_t, VL_t\}$
- $\mathbf{M} = \{M_{VL_t}, M_{Weight_t}, M_{EFV_t}, M_{MEMS_t}\}$
- $\mathbf{Z} = \{TI_t, MV_t, BMQ, SES, BHV\}$.

Note that under $\mathbf{G}_{alt2}$ ($\mathbf{G}_{main}$ with additional arrows SES $\rightarrow$ MEMS$_t$), BHV becomes a part of the c-DAG.

4.3. Identifiability and recoverability of causal effects

To make a decision on the recoverability of a causal query of interest, we first need to find out whether the query is identifiable or not. Provided the identifiability holds true, the main idea of recoverability is to transform the partially observed variables from the identified expression into observables with the help of d-separation statements resulting from the m-DAG.

In this work, we aim to estimate the causal impact of the efavirenz drug concentration in the plasma of children with HIV, measured and controlled over a specific time period, on viral failure.

Particularly, we focus on two causal effects,

$$\theta_{36} := P(VL_{36} = 1;\ do(EFV_6 := a, EFV_{36} := a)) \tag{1}$$

and

$$\theta_{84} := P(VL_{84} = 1; do(EFV_6 := a, EFV_{36} := a, EFV_{48} := a, EFV_{60} := a, EFV_{84} := a)), \tag{2}$$

corresponding to the probability of viral failure after 36 and 84 wks under a fixed intervention on the plasma concentration of efavirenz drug at each previous up to current study time point. Note that for assessment of recoverability of the causal effects, only variables from the previous up to current study time point are considered, and we therefore focus on the corresponding subgraph of the "full" m-DAG in Fig. 3 containing the variables up to and including the 36th study week when deciding about identifiability and recoverability of θ_{36}.

We first consider the m-DAG $\mathbf{G}_{main}$ from Fig. 3, ignoring the dashed blue lines from VL_t to MV_t, $t \in \{0, 6, 36, 48, 60, 84\}$. Note that according to the taxonomy of the missingness mechanisms proposed by Mohan et al. (2013), the missingness is of MNAR type because TI_t, $t \in \{0, 6, 36, 48, 60, 84\}$, are unobserved variables directly causing missingness. Under Rubin's definition the data would also be MNAR because units may exhibit missing values due to unmeasured behavioral factors (BHV).

In the first step, we have to work on identifiability of the causal effect of interest. We carry this out through two approaches. Our first proposal is the application of the IDC algorithm. Applying
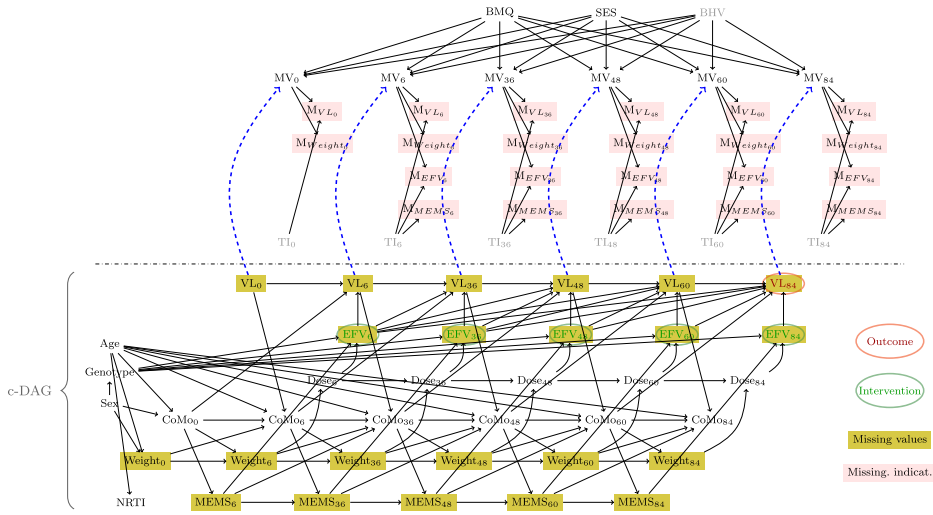
Figure 3: m-DAG. Dashed arrows distinguish between $\mathbf{G}_{main}$ (no dashed arrows) and $\mathbf{G}_{alt1}$ (dashed arrows). The subgraph below the dashed line represents the corresponding c-DAG. Unmeasured variables are slightly shaded. Variable names: Sex: biological sex, Weight: body weight, Age: age (in yrs), VL: elevated viral load, Dose: efavirenz dose administered, EFV: efavirenz concentration (12h after dose), NRTI: nucleoside reverse transcriptase inhibitors drug, Genotype: metabolism status, MEMS: adherence to treatment (measured through memory caps), CoMo: co-morbidities, MV: missed hospital visit, TI: technical issues (e.g. with blood samples or memory caps), BMQ: beliefs and attitudes toward medicine, SES: socio-economic status of caregiver, BHV: behavioral factors.

rules of *do-calculus* therefore leads to the causal effect identifiable in terms of the observed data distribution as if there were no missingness in any variable relevant for identification:

$$
\begin{aligned}
\theta_{36} = \sum_{\substack{Age, Sex, CoMo_0, Weight_0, \\ VL_0, CoMo_6, VL_6}} & P(VL_{36} = 1 | Age, Sex, CoMo_0, Weight_0, Genotype, MEMS_6, Weight_6, Dose_6, \\
& EFV_6 = a, VL_0, CoMo_6, VL_6, MEMS_{36}, Weight_{36}, Dose_{36}, EFV_{36} = a) \\
& P(VL_6 | Age, Sex, CoMo_0, Weight_0, Genotype, MEMS_6, Weight_6, \\
& Dose_6, EFV_6 = a, VL_0) \\
& P(CoMo_6 | Age, Sex, CoMo_0, Weight_0, Genotype, VL_0) \\
& P(Weight_0 | Age, Sex) P(CoMo_0 | Age, Sex) \\
& P(Sex) P(Age) P(VL_0).
\end{aligned}
$$

$$(3)$$

The corresponding identifiability result for θ_{84} based on IDC algorithm is in Appendix B.1.

There may be other identifiability results, based on different factorizations. For example, in our second proposal, we apply Pearl's (generalized) back door-criterion to determine a sufficient adjustment set and then simply apply Robins' *g*-formula factorization to it. Such a factorization may involve more conditional distributions that have to be estimated compared to the first approach.

For the above estimand, θ_{36}, a valid factorization would be as follows (see, e.g. Hernán and Robins (2020a, Chapters 19 and 21)):

$$\theta_{36} = \sum_{\substack{Age, Genotype, Sex, NRTI, Weight_0, VL_0 \\ Weight_6, MEMS_6, Weight_{36}, MEMS_{36}}} P(VL_{36} = 1 | EFV_6 = a, EFV_{36} = a, Age, Genotype, Sex, NRTI,$$

$$Weight_0, Weight_6, MEMS_6, Weight_{36}, MEMS_{36}, VL_0, VL_6)$$

$$P(Weight_{36} | EFV_6 = a, Age, Genotype, Sex, NRTI,$$

$$VL_0, VL_6, Weight_0, Weight_6, MEMS_6) \qquad (4)$$

$$P(MEMS_{36} | EFV_6 = a, Age, Genotype, Sex, NRTI,$$

$$VL_0, VL_6, Weight_0, Weight_6, MEMS_6)$$

$$P(VL_6 | EFV_6 = a, Age, Genotype, Sex, NRTI, VL_0, Weight_0)$$

$$P(Weight_6, MEMS_6, Age, Genotype, Sex, NRTI, VL_0, Weight_0)$$

This is because neither the dose nor co-morbidities variables are necessarily required to block the relevant back-door paths from EFV_t to VL_{t^*}, $t^* \geq t$, i.e. those back-door paths that do not go through any future concentrations.

Note that the identified expressions for θ_{36} and θ_{84} are identical under $\mathbf{G}_{main}$ and $\mathbf{G}_{alt1}$ because the c-DAG is identical in both situations. Under assumption of $\mathbf{G}_{alt2}$, the c-DAG changes by addition of the variable BHV, resulting in a slight change of the identified expression (see Appendix C.2).

In the second step, we need to decide on the recoverability of the identified expression in terms of the observed data. Note that the missingness mechanism depicted in $\mathbf{G}_{main}$ refers to the *closed missingness mechanism* introduced in the previous section. In this case, there is no path between any variable from the c-DAG containing the variables of interest and the set of missingness indicators M and their causes. Therefore, both causal effects are recoverable, and an available case analysis is admissible in this situation. In order to provide a better intuition about how to decide on the recoverability of a causal effect, we present the recoverability result for θ_{36} in Appendix C.1, implicitly referring to a situation for which no recoverability result exists, and it has to be decided on recoverability manually.

If the m-DAG $\mathbf{G}_{main}$ reflects relationships in the data correctly, then an available case analysis is sufficient for estimating the causal effects of interest consistently. However, it is also necessary to assess the plausibility of the graph structure assumption. A suggestion would be to perform a sensitivity analysis to investigate the recoverability of the causal effects θ_{36} and θ_{84} under the assumption of an alternative m-DAG which is eligible in terms of content. Therefore, we consider another m-DAG, $\mathbf{G}_{alt1}$, which is depicted in Fig. 3 when including the dashed blue lines from VL_t to MV_t, $t \in \{0, 6, 36, 48, 60, 84\}$. The arrows from VL_t (binary variable *elevated viral load*) to MV_t (binary variable *missed visit*) for the respective measurement time point reflect the speculation that children with viral failure may miss their appointments due to their poor health condition.

We first investigate the recoverability of θ_{36} under the assumption of $\mathbf{G}_{alt1}$ being the true underlying m-DAG. Note that if at least one of the conditional distributions in Equation 3 is non-recoverable, we conclude the non-recoverability of the causal query of interest.

We first focus on the second last conditional distribution in Equation 3 which involves the partially observed data, $P(VL_6 | EFV_6 = a, Age, Genotype, Sex, NRTI, VL_0, Weight_0)$, and the goal is to condition on $M_{Weight_0} = 0$, $M_{VL_0} = 0$, $M_{EFV_6} = 0$ and $M_{VL_6} = 0$. From the m-DAG (assuming faithfulness), we know that $VL_6 \perp\!\!\!\perp (M_{Weight_0}, M_{VL_0}, M_{EFV_6}, M_{VL_6}) | (MV_0, MV_6)$ (and we always need to condition on (at least) MV_0 and MV_6 to achieve (conditional) independence), which in

turn leads to

$$P(VL_6^{M=0}|EFV_6^{M=0} = a, Age, Genotype, Sex, NRTI, VL_0^{M=0}, Weight_0^{M=0})$$

$$= \sum_{MV_0, MV_6} P(VL_6^{M=0}|EFV_6^{M=0} = a, Age, Genotype, Sex, NRTI, VL_0^{M=0}, Weight_0^{M=0}) \times$$

$$P(MV_0, MV_6|EFV_6^{M=0} = a, Age, Genotype, Sex, NRTI, VL_0^{M=0}, Weight_0^{M=0})$$

$$= \sum_{MV_0, MV_6} P(VL_6^{M=0}|EFV_6^{M=0} = a, Age, Genotype, Sex, NRTI, VL_0^{M=0}, Weight_0^{M=0},$$

$$M_{Weight_0} = 0, M_{VL_0} = 0, M_{EFV_6} = 0, M_{VL_6} = 0) \times$$

$$P(MV_0, MV_6|EFV_6 = a^{M=0}, Age, Genotype, Sex, NRTI, VL_0^{M=0}, Weight_0^{M=0})$$

$$= \sum_{MV_0, MV_6} P(VL_6|EFV_6 = a, Age, Genotype, Sex, NRTI, VL_0, Weight_0,$$

$$M_{Weight_0} = 0, M_{VL_0} = 0, M_{EFV_6} = 0, M_{VL_6} = 0) \times$$

$$P(MV_0, MV_6|EFV_6 = a^{M=0}, Age, Genotype, Sex, NRTI, VL_0^{M=0}, Weight_0^{M=0}).$$

The second factor in the expression above,

$$P(MV_0, MV_6|EFV_6^{M=0} = a, Age, Genotype, Sex, NRTI, VL_0^{M=0}, Weight_0^{M=0}),$$

cannot be further decomposed following the adjustment formula, as among others we would need MV_6 to be conditionally independent of M_{EFV_6} (possibly given some other variables), but there is no such subset because MV_6 is a parent of M_{EFV_6}.

Because other available recoverability techniques, like sequential factorization or R factorization (Mohan and Pearl 2021), are not applicable in our case, we conclude that the conditional distribution cannot be expressed in terms of the observed data distribution only; this results in non-recoverability of θ_{36}. Based on the same arguments, non-recoverability of θ_{84} also follows.

4.4. Analysis

The analysis is conducted using the m-DAG $\mathbf{G}_{main}$ from Fig. 3 and the study data described in Section 4.1. The plug-in g-formula estimation [see, e.g. Hernan and Robins (Hernán and Robins 2020a, Chapter 13)] of our estimands 1 and 2 was based on equations 4 and 6, respectively. Given the recoverability under $\mathbf{G}_{main}$, we conduct the analysis based on available cases. For modeling the respective conditional distributions, between 69 and 85 complete observations are available.

Note that in our setting, relevant co-morbidities in the DAG refer to AIDS-defining events, including tuberculosis, persistent diarrhea, malnutrition, and severe wasting, among others. Most of those events were not measured regularly in our study and thus were not included in the analysis. However, those co-morbidities are not expected to be very frequent in the study population. This is because children with active infections, those being treated for tuberculosis and with severe laboratory abnormalities at screening, were not enrolled into the study, and most children were in relatively good health (Mulenga et al. 2016). Also, not all valid identification formulae require information on them, such as those used in our analysis. As adherence could not be measured regularly, as explained in Section 4.1, we constructed a variable indicating any signs of non-adherence, defined as the mean memory caps opening percentage being less than 90%.

Although the main analysis was based on available cases due to the identification results, we also implemented a multiple imputation analysis. We consider the results of the multiple imputation analysis invalid because the underlying missingness mechanism is assumed to be MNAR. This is
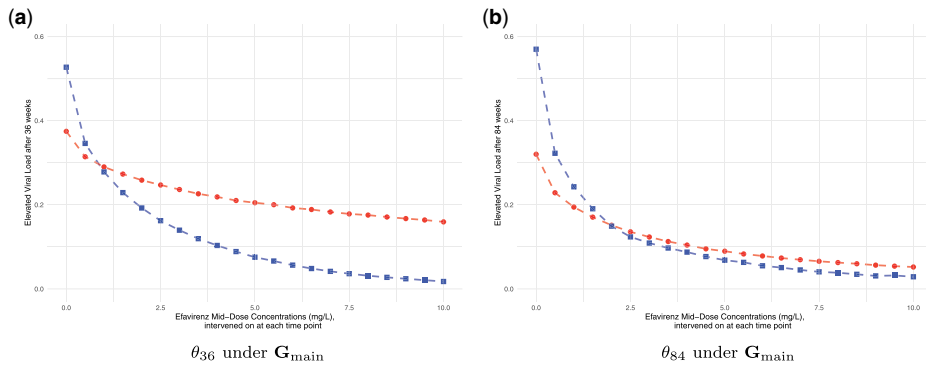
Figure 4: Estimated CCRCs for θ_{36} and θ_{84} under m-DAG $\mathbf{G}_{\mathrm{main}}$. (a) Estimated CCRCs for θ_{36}. (b) Estimated CCRCs for θ_{84}. Causal effects were estimated using available cases (blue squares) and multiple imputation, mean over 20 imputed data sets (red dots); results represent the mean over 1000 seeds.

primarily due to the unobserved variables (TI_t), which directly cause missingness in variables from the c-DAG.

We multiply imputed five data sets using the Expectation-Maximization Bootstrap (EMB) Algorithm, a joint modeling-based imputation approach implemented in the *R*-package `Amelia II`. Our setup considered the longitudinal setup through lag-variables and the inclusion of splines of time. In addition, to improve numerical stability of the EM algorithm, we added a ridge prior (which shrinks the covariances among the variables toward zero without changing the means or variances). This is often recommended when using EMB on small samples (Honaker et al. 2011). To address the fact that age, weight and EFV concentrations can not have negative or other illogical values, we specified lower and upper bounds for those variables in the EMB algorithm: 0 and 35 for EFV, 0 and 3 for log age and 2 and 4 for log weight. `Amelia II` implements these bounds by rejection sampling: when drawing the imputations from the respective posterior distributions, all logical constraints need to be satisfied; otherwise, imputations are redrawn until those constraints are met.

The results of our analyses are presented in Fig. 4a and b.

One can see a higher probability of viral failure with lower EFV concentration values, independent of the approach used to address the missing data. Using the suggested available cases approach, the probability of failure is estimated to be $> 50\%$ at both 36 and 84 wks if concentrations were 0mg/L, e.g. if patients did not take any medications. With higher concentrations, failure probabilities decrease to below 5%, which is expected as EFV is expected to be a potent drug. Interestingly, the CCRCs estimated using multiple imputation are much flatter than under the available case approach, both for θ_{36} and θ_{84}.

We will now explore the derived theoretical results, along with comparisons between the available case and multiple imputation approaches in the simulation studies below.

5. SIMULATION STUDIES

In this section, through simulation studies, we assess the reliability of CCRC estimates for θ_{84} in the presence of missing data. We compare true CCRC values with estimates derived from: (i) an ideal scenario with no missing data (complete data analysis), (ii) available case analysis, and (iii) multiple imputation.

5.1. Setup

The data were simulated using the R-package `simcausal` (Sofrygin et al. 2017), which is a powerful and flexible tool for simulation of longitudinal data structures based on structural causal models. A structural causal model (SCM) uniquely imposes a causal graph, allowing data structures to be generated according to such a causal DAG (in our case, an m-DAG).

Simulation 1: We simulate data corresponding to both m-DAGs, $\mathbf{G}_{main}$ and $\mathbf{G}_{alt1}$, as defined in Fig. 3. We first simulate binary and normally distributed confounder sets, a continuous (truncated normally distributed) intervention and a binary outcome for all 6 time points and for the sample size of $n = 5.000$. This way, the data aligning with the c-DAG is simulated. Afterwards, we simulate the binary variables causing missingness (SES_t (as proxy for BMQ_t, SES_t and BHV_t), MV_t and TI_t, $t \in \{0, 6, 36, 48, 60, 84\}$. Based on these variables, we simulate the missingness indicators for VL_t, $t \in \{0, 6, 36, 48, 60, 84\}$, and $MEMS_t$, $t \in \{6, 36, 48, 60, 84\}$. In this scenario, we assume that only these two variables (for all time points) are subject to missing data.

Simulation 2: The DGP is the same as for Simulation 1, but we simulate missingness indicators for many more variables: VL_t, $Weight_t$, $t \in \{0, 6, 36, 48, 60, 84\}$, and $MEMS_t$, EFV_t, $t \in \{6, 36, 48, 60, 84\}$. This scenario mostly coincides with the GPD induced by Fig. 3. This scenario introduces another layer of complexity because the distribution of EFV_t, $t \in \{6, 36, 48, 60, 84\}$ is non-symmetric and complex. Using a parametric imputation model with slightly misspecified distributional assumptions may introduce some bias. This setting serves as a benchmark for a realistic scenario where missing data with somewhat complex distributions are imputed using parametric assumptions in conjunction with predictive mean matching (fully conditional imputation procedures, like (M)ICE), or variations thereof (joint modeling, as in `Amelia II` which we use below).

The exact model specifications are given in Appendix D.1. Based on the DGP, we first simulate the interventional data (1000 repetitions) and compute the true CCRC based on it. We intervene on EFV at time points 6, 36, 48, 60, and 84, considering a sequence of interventions from 0 up to 10 mg/L with steps of 0.5 mg/L. Next, the CCRC is estimated using the g-formula factorization provided in Appendix B.1, once on the complete data and once on the available case data. Compared to the widely used complete case analysis, which relies on samples where all variables in the data set are observed, the available case analysis retains all samples where variables in the query of interest (for the g-formula, the variables required for the estimation of the conditional distribution of interest) are observed. For this reason, an available case analysis is preferred over a complete case analysis due to a more economical usage of available data (Mohan and Pearl 2021). Finally, the CCRC is estimated using multiple imputation (MI) and the g-formula.

5.2. Results

Below, we focus on the results from Simulation 1. The results of the Simulation 2 are presented in Appendix D.2, specifically in Figs. D.1a and D.1b.

The simulation results align with the theoretical findings. We first consider the estimated CCRC curve for elevated viral load after 84 wks based on data simulated under $\mathbf{G}_{main}$, as presented in Fig. 5a. According to the findings in Section 4, the missingness mechanism is of MNAR type. Despite this, the causal query of interest, θ_{84}, remains recoverable, allowing for consistent estimation of the CCRC via an available case analysis. Conversely, the MI approach is inadmissible due to its underlying assumption of Missing At Random (MAR), which is not met because of the partially observed variables TI_t, $t \in 0, 6, 36, 48, 60, 84$, that affect the probability of missingness in variables from $\mathbf{V}_m$. The estimated counterfactual outcomes, illustrated in Fig. 5b, represent the mean estimated values across 1000 simulation repetitions. It is evident that the CCRC estimates from available case analysis (dashed blue line) match the true CCRC (dashed green line), whereas those derived from MI differ markedly. To determine whether the discrepancies represent a bias or could also be explained by simulation uncertainty, we calculated Monte Carlo confidence intervals for the differences between the MI estimation results and the true causal effects θ_{84}.
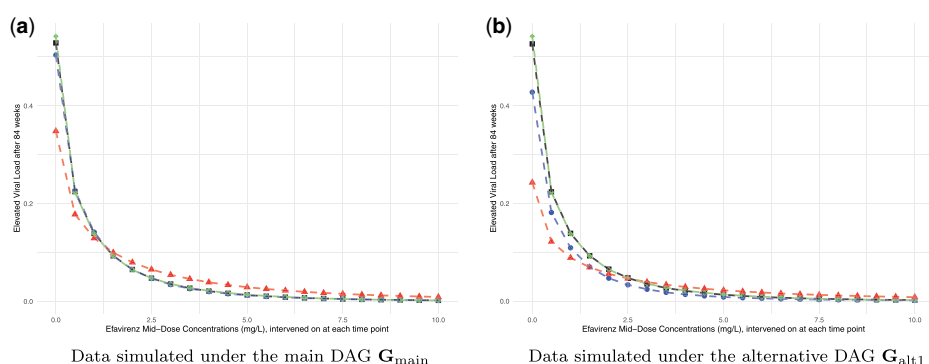
Figure 5: Estimated CCRCs for the probability of viral failure after 84 wks. (a) Data simulated under the main DAG $\mathbf{G}_{\text{main}}$. (b) Data simulated under the alternative DAG $\mathbf{G}_{\text{alt1}}$. Causal effects were estimated on complete data (black squares), incomplete data using available cases (blue dots), incomplete data using multiple imputation (red triangles) and counterfactual data (green diamonds, true CCRC); results represent the mean over 1000 seeds.

The results are reported in Table D.1a and show that for each EFV value, the estimated differences are significantly different from zero. This discrepancy underscores the theoretical findings regarding the bias introduced by MI due to the MNAR missingness mechanism.

Secondly, consider the results under the alternative DAG $\mathbf{G}_{alt1}$, presented in Fig. 5b. In this case, the causal query of interest is non-recoverable. The results confirm this and demonstrate that neither the available case analysis (dashed blue line) nor the MI estimates (dashed red line) align with the true CCRC (dashed green line). The respective Monte Carlo confidence intervals, see Tables D.1b and D.1c in the appendix, show that these differences cannot be explained by simulation uncertainty alone, which is consistent with the theoretical findings.

These findings highlight the critical need to scrutinize the common MAR-type missingness assumption, especially in complex longitudinal data scenarios where multiple variables experience missingness. Our simulation study illustrates how even minor changes in the missingness structure can dramatically affect recoverability and, consequently, the accuracy of estimation results. This emphasizes the importance of careful consideration and justification of missingness assumptions in such analyses.

6. CONCLUSIONS

Our analyses demonstrate the applicability of missingness DAGs to complex longitudinal studies and show that, in some cases, available case analyses can be valid under MNAR. However, our application also highlights the substantial effort involved, the technical expertise required, and sensitivity of results to assumptions in the causal model.

Under the assumptions represented in Fig. 3, the data are missing not at random (MNAR) because technical issues (TI) with pill containers (frequently) and blood samples (rarely), which we assume to be direct causes of missingness in multiple variables, have not been measured. However, we show that these assumptions are sufficient to estimate the desired impact using the available data and g-formula representations, despite the MNAR mechanism. However, if reasons for missed visits are caused by the outcome (elevated viral load), as speculated in the alternative DAG, the causal effect cannot be recovered. Interestingly, additional simulations (Appendix D.2, Fig. D.2) show that recoverability holds true even if behavioral factors directly cause non-adherence. This corresponds to the situation discussed in the second alternative DAG, G_{alt2}. However, it is possible that more complex processes exist between socio-economic/behavioral factors and biologic processes, for which identifiability does not hold. Our investigations demonstrate the sensitivity of recoverability

results to even minor changes in the missingness structure. We therefore emphasize the need for careful inspection of the assumptions regarding the missing data mechanism, especially in complex longitudinal studies with multiple variables experiencing missingness.

It is also important to highlight that the estimated concentration-response curves are much flatter when using multiple imputation and are actually invalid under the assumptions stated in Fig. 3, as predicted by theory and confirmed by our simulations.

Our analyses show lower probabilities of viral failure with higher concentrations after accounting for the missing data. There are, of course, many further complications that may have to be considered for accurate causal inference in our setting. First, one may also account for measurement error, e.g. in viral load and EFV concentrations (Schomaker et al. 2015), though this may not affect our binary viral failure definition very much. Second, it would be advantageous if measurements were available more frequently and precisely, especially for measuring actual adherence to the prescribed treatment plan, which is difficult in practice.

While advancements in causal modeling and appropriate statistical estimation techniques are impressive, answering complex epidemiological and biological questions remains a challenge.

A. STRUCTURAL EQUATION MODEL

We now present the non-parametric structural equation models (SEMs) corresponding to the m-DAGs in Fig. 3. Note that the follow-up time points 1 through 5 coincide with the study weeks 6, 36, 48, 60, and 84. The independent (joint) noise term is denoted as $\mathbf{U}$.

The SEM corresponding to $\mathbf{G}_{main}$ in Fig. 3 is as follows:

For $t = 0$:

$$\text{Genotype} = f_{\text{Genotype}}(\text{Sex}, U_{\text{Genotype}})$$
$$\text{Weight}_0 = f_{\text{Weight}_0}(\text{Sex}, \text{Age}, U_{\text{Weight}_0})$$
$$\text{NRTI}_0 = f_{\text{NRTI}}(\text{Age}, U_{\text{NRTI}})$$
$$\text{MV}_0 = f_{\text{MV}_0}(\text{BMQ}, \text{SES}, \text{BHV}, U_{\text{MV}_0})$$

For $t = 1$:

$$\text{Dose}_1 = f_{\text{Dose}_1}(\text{Weight}_1, U_{\text{Dose}_1})$$

For $t \geq 0$:

$$\text{M}_{Weight_t} = f_{\text{M}_{Weight_t}}(\text{MV}_t, U_{\text{M}_{Weight_t}})$$
$$\text{M}_{VL_t} = f_{\text{M}_{VL_t}}(\text{MV}_t, \text{TI}_t, U_{\text{M}_{VL_t}})$$

For $t \geq 1$:

$$\text{MEMS}_t = f_{\text{MEMS}_t}(\text{CoMo}_{t-1}, \text{MEMS}_{t-1}, U_{\text{MEMS}_t}) \quad [\text{assume MEMS}_0 = 0]$$
$$\text{Weight}_t = f_{\text{Weight}_t}(\text{Weight}_{t-1}, \text{CoMo}_{t-1}, U_{\text{Weight}_t})$$
$$\text{CoMo}_t = f_{\text{CoMo}_t}(\text{CoMo}_{t-1}, \text{Age}, \text{Weight}_{t-1}, \text{VL}_{t-1}, U_{\text{CoMo}_t})$$
$$\text{EFV}_t = f_{\text{EFV}_t}(\text{Dose}_t, \text{MEMS}_t, \text{Genotype}, U_{\text{EFV}_t})$$
$$\text{VL}_t = f_{\text{VL}_t}(\text{VL}_{t-1}, \text{CoMo}_{t-1}, \text{EFV}_t, U_{\text{VL}_t})$$
$$\text{MV}_t = f_{\text{MV}_t}(\text{BMQ}, \text{SES}, \text{BHV}, U_{\text{MV}_t})$$
$$\text{M}_{EFV_t} = f_{\text{M}_{EFV_t}}(\text{MV}_t, \text{TI}_t, U_{\text{M}_{EFV_t}})$$
$$\text{M}_{MEMS_t} = f_{\text{M}_{MEMS_t}}(\text{TI}_t, U_{\text{M}_{MEMS_t}})$$

For $t \geq 2$:

$$\text{Dose}_t = f_{\text{Dose}_t}(\text{Dose}_{t-1}, \text{Weight}_t, U_{\text{Dose}_t})$$

The SEM for $\mathbf{G}_{alt1}$ (with present blue dashed lines in Fig. 3) is the same as the SEM above, except for the structural equations for MV_t, $t \in \{0, 6, 36, 48, 60, 84\}$. These are specified as follows for $\mathbf{G}_{alt1}$:

For $t = 0$:

$$\text{MV}_0 = f_{\text{MV}_0}(\text{BMQ, SES, BHV, VL}_0, U_{\text{MV}_0})$$

For $t \geq 1$:

$$\text{MV}_t = f_{\text{MV}_t}(\text{BMQ, SES, BHV, VL}_t, U_{\text{MV}_t})$$

This way, MV_t additionally depends on VL_t, $t \in \{0, 6, 36, 48, 60, 84\}$, which corresponds to the dashed blue lines in the DAG.

B. IDENTIFIABILITY RESULTS
B.1. Identifiability results for θ_{84}

The identifiability result below is based on application of the IDC algorithm (Shpitser and Pearl 2006).

$$\theta_{84} = \sum_{\substack{Age, Sex, CoMo_0, Weight_0, \\ MEMS_6, Weight_6, VL_0, CoMo_6, \\ VL_6, MEMS_{36}, Weight_{36}, CoMo_{36}, \\ MEMS_{48}, Weight_{48}, VL_{36}, CoMo_{48}, \\ VL_{48}, CoMo_{60}, VL_{60}}} P(VL_{84} = 1 | Age, Sex, CoMo_0, Weight_0, Genotype, MEMS_6, Weight_6,$$

$$Dose_6, EFV_6 = a, VL_0, CoMo_6, VL_6, MEMS_{36},$$

$$Weight_{36}, CoMo_{36}, Dose_{36}, MEMS_{48}, Weight_{48}, EFV_{36} = a,$$

$$Dose_{48}, VL_{36}, EFV_{48} = a, CoMo_{48}, VL_{48}, MEMS_{60}, Weight_{60}, CoMo_{60},$$

$$Dose_{60}, MEMS_{84}, Weight_{84}, EFV_{60} = a, Dose_{84}, VL_{60}, EFV_{84} = a)$$

$$P(VL_{60} | Age, Sex, CoMo_0, Weight_0, Genotype, MEMS_6, Weight_6,$$

$$Dose_6, EFV_6 = a, VL_0, CoMo_6, VL_6, MEMS_{36},$$

$$Weight_{36}, CoMo_{36}, Dose_{36}, MEMS_{48}, Weight_{48}, EFV_{36} = a, Dose_{48},$$

$$VL_{36}, EFV_{48} = a, CoMo_{48}, VL_{48}, MEMS_{60}, Weight_{60}, Dose_{60}, EFV_{60} = a)$$

$$P(CoMo_{60} | Age, Sex, CoMo_0, Weight_0, Genotype, MEMS_6, Weight_6,$$

$$Dose_{36}, Dose_6, EFV_6 = a, VL_0, CoMo_6, VL_6, MEMS_{36}, Weight_{36}, CoMo_{36},$$

$$MEMS_{48}, Weight_{48}, EFV_{36} = a, Dose_{48}, VL_{36}, EFV_{48} = a, CoMo_{48}, VL_{48})$$

$$P(VL_{48} | Age, Sex, CoMo_0, Weight_0, Genotype, MEMS_6, Weight_6, Dose_6,$$

$$EFV_6 = a, VL_0, CoMo_6, VL_6, MEMS_{36}, Weight_{36}, CoMo_{36}, Dose_{36},$$

$$MEMS_{48}, Weight_{48}, EFV_{36} = a, Dose_{48}, VL_{36}, EFV_{48} = a)$$

$$P(CoMo_{48} | Age, Sex, CoMo_0, Weight_0, Genotype, MEMS_6, Weight_6,$$

$$Dose_6, EFV_6 = a, VL_0, CoMo_6, VL_6, MEMS_{36},$$

$$Weight_{36}, CoMo_{36}, Dose_{36}, EFV_{36} = a, VL_{36})$$

$$P(Weight_{48}|Age, Sex, CoMo_0, Weight_0, Genotype, MEMS_6, Weight_6,$$
$$Dose_6, EFV_6 = a, VL_0, CoMo_6, VL_6, Weight_{36}, CoMo_{36})$$

$$P(MEMS_{48}|Age, Sex, CoMo_0, Weight_0, Genotype, MEMS_6, Weight_6,$$
$$Dose_6, EFV_6 = a, VL_0, CoMo_6, VL_6, MEMS_{36}, CoMo_{36})$$

$$P(CoMo_{36}|Age, Sex, CoMo_0, Weight_0, Genotype, MEMS_6, Weight_6,$$
$$Dose_6, EFV_6 = a, VL_0, CoMo_6, VL_6)$$

$$P(VL_{36}|Age, Sex, CoMo_0, Weight_0, Genotype, MEMS_6, Weight_6, Dose_6,$$
$$EFV_6 = a, VL_0, CoMo_6, VL_6, MEMS_{36}, Weight_{36}, Dose_{36}, EFV_{36} = a)$$

$$P(Weight_{36}|Age, Sex, CoMo_0, Weight_0, Genotype, Weight_6,$$
$$VL_0, CoMo_6)P(MEMS_{36}|Age, Sex, CoMo_0, Weight_0, Genotype, MEMS_6,$$
$$VL_0, CoMo_6)P(VL_6|Age, Sex, CoMo_0, Weight_0, Genotype,$$
$$MEMS_6, Weight_6, Dose_6, EFV_6 = a, VL_0)$$

$$P(CoMo_6|Age, Sex, CoMo_0, Weight_0, Genotype, VL_0)$$

$$P(Weight_6|Age, Sex, CoMo_0, Weight_0)P(MEMS_6|CoMo_0)$$

$$P(Weight_0|Age, Sex)P(CoMo_0|Age, Sex)P(VL_0)P(Sex)P(Age) \tag{5}$$

An alternative identifiability result can be obtained, for example, by applying the generalized back-door criterion. This results in using an adjustment set given by the confounders weight and adherence (MEMS). Both dose and co-morbidities are not necessarily required to block the relevant back-door paths from EFV_t to VL_{t^*}, $t^* \geq t$, i.e. the back-door paths that do not pass through any future concentrations (Hernán and Robins 2020a). Applying Robins' parametric g-formula (Robins 1986; Hernán and Robins 2020a) leads to the following valid g-formula factorization:

$$\theta_{84} = \int_l \left\{ P(VL_{84} = 1|\bar{A}_{84} = \bar{a}_{84}, \bar{\mathbf{L}}_{84} = \bar{\mathbf{l}}_{84}) \right.$$
$$\left. \times \prod_{s=0}^{84} f(\mathbf{L}_s = \mathbf{l}_s|\bar{A}_{s-1} = \bar{a}_{s-1}, \bar{\mathbf{L}}_{s-1} = \bar{\mathbf{l}}_{s-1}) \right\} dF_l(l), \tag{6}$$

where $F_L(\cdot)$ denotes the CDF with respect to L. In the above, $\mathbf{L}_t = \{weight_t, MEMS_t\}$, $\mathbf{L}_0 = \{weight_0, NRTI, Genotype, Sex, Age\}$ and $A_t = EFV_t$. Note that the overbar refers to the intervention and confounder histories, i.e. for of a unit i (up to and including time t) the histories are $\bar{A}_{t,i} = (A_{0,i}, \ldots, A_{t,i})$ and $\bar{L}_{t,i}^s = (L_{0,i}^s, \ldots, L_{t,i}^s)$, $s = 1, 2$, $i = 1, \ldots, n$, respectively. Outcomes (i.e. VL) prior to s are part of $\bar{\mathbf{L}}_s$. The inner product of the 2 time-varying and ordered confounders $\mathbf{L}_s = \{L_s^1, L_s^2\}$ can be decomposed further:

$$\prod_{s=0}^{84} f(L_s^2 = l_s^2 \mid L_s^1 = l_s^1, \bar{A}_{s-1} = \bar{a}_{s-1}, \bar{\mathbf{L}}_{s-1} = \bar{\mathbf{l}}_{s-1}) \times f(L_s^1 = l_s^1 \mid \bar{A}_{s-1} = \bar{a}_{s-1}, \bar{\mathbf{L}}_{s-1} = \bar{\mathbf{l}}_{s-1})$$

Thus, implementing (6), requires fitting models for the outcome and two confounder distributions at each time point, given their respective history.

C. RECOVERABILITY RESULTS FOR θ_{36}

C.1. Assuming G_{main}

To assess the recoverability of the identified causal effect θ_{36} in Equation 3, we must show that each of the multiplicative factors corresponding to the conditional or marginal distributions can be expressed in terms of the observed data only. Although the missingness mechanism is of the "closed" type, and the recoverability result is directly available for this case, we want to demonstrate how one would examine identifiability if such a result is not available. T do so, we present the result for one of the conditional distributions containing partially observed variables. The results for other multiplicative factors are derived analogously.

$$P(CoMo_6^{M=0}|Age, Sex, CoMo_0^{M=0}, Weight_0^{M=0}, Genotype, VL_0^{M=0})$$
$$= P(CoMo_6^{M=0}|Age, Sex, CoMo_0^{M=0}, Weight_0^{M=0}, Genotype, VL_0^{M=0}, \mathbf{M} = 0)$$
$$= P(CoMo_6|Age, Sex, CoMo_0, Weight_0, Genotype, VL_0, \mathbf{M} = 0)$$

The first equality holds due to the fact that all partially observed variables are independent of the corresponding relevant missingness indicators, whereas the second equality is true due to the consistency assumption (Moreno-Betancur et al. 2018).

Note that the recoverability result above is trivial due to the specific missingness mechanism. In a general case, much more effort is required to recover a distribution of interest.

C.2. Assuming G_{alt2}

We consider a second plausible alternative m-DAG, G_{alt2}, which is equivalent to G_{main} in Fig. 3, with the addition of a direct effect of BHV on $MEMS$ for weeks 6, 36, 48, 60 and 84. In this model, we propose that behavioral pattern may affect adherence.

Under G_{alt2}, the identified expression for the causal effect of interest, θ_{36}, is as follows:

$$
\begin{aligned}
\theta_{36} = \sum_{\substack{Age, Sex, CoMo_0, Weight_0, \\ VL_0, CoMo_6, VL_6}} & P(VL_{36}|Age, Sex, BHV, CoMo_0, Weight_0, Genotype, MEMS_6, \\
& Weight_6, Dose_6, EFV_6 = a, VL_0, CoMo_6, VL_6, MEMS_{36}, Weight_{36}, \\
& Dose_{36}, EFV_{36} = a) \\
& P(VL_6|Age, Sex, CoMo_0, Weight_0, Genotype, MEMS_6, \\
& Weight_6, Dose_6, EFV_6 = a, VL_0) \\
& P(CoMo_6|Age, Sex, CoMo_0, Weight_0, Genotype, VL_0) \\
& P(Weight_0|Age, Sex)P(CoMo_0|Age, Sex)P(VL_0)P(Sex)P(Age).
\end{aligned}
$$

(7)

Note that even if the identified expression looks almost identical to the one we obtain under G_{main} or G_{alt1} (compare Equation 3), the recoverability result may differ due to the different conditional independence statements that hold in G_{alt2}, assuming it is faithful.

D. SIMULATION STUDIES

D.1. DGP for simulations 1 and 2

Both baseline data $(t = 0)$ and follow-up data $(t = 1, \ldots, 5)$ were created using structural equations with the R-package simcausal. Note that the follow-up time points 1 through 5 correspond to the study weeks 6, 36, 48, 60 and 84. The distributions listed below, in temporal order, describe the data-generating process. Our baseline data consists of *Sex, Genotype*, log(age) (*Age*), log(weight) (*Weight*), the respective Nucleoside Reverse Transcriptase Inhibitor (*NRTI*), and a proxy for

socio-economic status (SES). The time-varying variables are co-morbidities (CoMo), efavirenz dose (Dose), efavirenz mid-dose concentration (EFV), elevated viral load (= viral failure, VL), adherence (measured through memory caps, MEMS), missed visit (MV), technical issues (TI), and the missingness indicators for *EFV*, *Weight*, *VL*, and *MEMS*, respectively. In addition to Bernoulli (B), Poisson (*Poisson*), Multinominal (*MN*) and Normal (N) distributions, we also use truncated normal distributions, denoted by $N_{[a,a_1,a_2,b,b_1,b_2]}$, where a and b are the truncation levels. Values smaller than a are replaced by a random draw from a $U(a_1, a_2)$ distribution and values greater than b are drawn from a $U(b_1, b_2)$ distribution, where U refers to a continuous uniform distribution. For the specified multinomial distributions, probabilities are normalized, if required, to ensure they add up to 1.

The DGP corresponding to $\mathbf{G}_{main}$ in Fig. 3 is as follows:

For $t = 0$:

$$\text{Sex}_0 \sim B(p = 0.5)$$

$$\text{Genotype}_0 \sim MN \begin{pmatrix} p1 = 1/(1 + \exp(-(-0.103 + I(\text{Sex}_0 = 1) \times 0.223 + I(\text{Sex}_0 = 0) \times 0.173))), \\ p2 = 1/(1 + \exp(-(-0.086 + I(\text{Sex}_0 = 1) \times 0.198 + I(\text{Sex}_0 = 0) \times 0.214))), \\ p3 = 1/(1 + \exp(-(-0.309 + I(\text{Sex}_0 = 1) \times 0.082 + I(\text{Sex}_0 = 0) \times 0.170))) \end{pmatrix}$$

$$\text{Age}_0 \sim N_{[0.693, 0.693, 1, 2.8, 2.7, 2.8]}(\mu = 1.501, \sigma = 0.369)$$

$$\text{Weight}_0 \sim N_{[2.26, 2.26, 2.67, 3.37, 3.02, 3.37]}(\mu = (1.5 + 0.2 \times \text{Sex} + 0.774 \times \text{Age}_0) \times 0.94, \sigma = 0.369)$$

$$\text{NRTI}_0 \sim MN \begin{pmatrix} p1 = 1/(1 + \exp(-(-0.006 + I(\text{Age}_0 > 1.4563) \times \text{Age}_0 \times 0.1735 + I(\text{Age}_0 \leq 1.4563) \times \text{Age}_0 \times 0.1570))), \\ p2 = 1/(1 + \exp(-(-0.006 + I(\text{Age}_0 > 1.4563) \times \text{Age}_0 \times 0.1735 + I(\text{Age}_0 \leq 1.4563) \times \text{Age}_0 \times 0.1570))), \\ p3 = 1/(1 + \exp(-(-0.006 + I(\text{Age}_0 > 1.4563) \times \text{Age}_0 \times 0.1570 + I(\text{Age}_0 \leq .14563) \times \text{Age}_0 \times 0.1818))) \end{pmatrix}$$

$$\text{CoMo}_0 \sim B(p = 0.15)$$

$$\text{VL}_0 \sim B(p = 1 - (1/(1 + \exp(-(0.4 + 1.9 \times \sqrt{\text{EFV}_0})))))$$

$$\text{SES}_0 \sim Poisson(\lambda = 3)$$

$$\text{MV}_0 \sim B(p = 1/(1 + \exp(-(-2.95 + 0.1 \times \text{SES}_0))))$$

For $t = 1$:

$$\text{Dose}_1 \sim MN \begin{pmatrix} p1 = 1/(1 + \exp(-(5 + \sqrt{(\text{Weight}_1)} \times 8 - \text{Age}_0 \times 10))), \\ p2 = 1/(1 + \exp(-(4 + \sqrt{(\text{Weight}_1)} \times 8.768 - \text{Age}_0 \times 9.06))), \\ p3 = 1/(1 + \exp(-(3 + \sqrt{(\text{Weight}_1)} \times 6.562 - \text{Age}_0 \times 8.325))), \\ p4 = 1 - (p1 + p2 + p3) \end{pmatrix}$$

For $t \geq 0$:

$$\text{TI}_t \sim B(p = 0.05)$$

$$M_{EFV_t} \sim B(p = I(\text{MV}_t = 1) + I(\text{MV}_t = 0) \times I(\text{TI}_t = 1) \times 0.5)$$

$$M_{Weight_t} \sim B(p = I(\text{MV}_t = 1))$$

$$M_{VL_t} \sim B(p = I(\text{MV}_t = 1) + I(\text{MV}_t = 0) \times I(\text{TI}_t = 1) \times 0.5)$$

For $t \geq 1$:

$$\text{MEMS}_t \sim B(p = 1/(1 + \exp(-(0.71 + \text{CoMo}_{t-1} \times 0.31 + \text{MEMS}_{t-1} \times I(t \geq 2) \times 0.31)))), \quad [\text{assume MEMS}_0 = 0]$$

$$\text{Weight}_t \sim N_{[2.26, 2.26, 2.473, 3.37, 3.2, 3.37]}(\mu = \text{Weight}_{t-1} \times 1.04 - 0.05 \times I(\text{CoMo}_{t-1} = 1), \sigma = 0.4)$$

$$\mathrm{CoMo}_t \sim \mathrm{B}(p = 1 - (1/(1 + \exp(-(0.5 \times \mathrm{I}(\mathrm{CoMo}_{t-1} = 1) + \mathrm{Age}_0 \times 0.1 + \mathrm{Weight}_{t-1} \times 0.1)))))$$

$$\mathrm{EFV}_t \sim \mathrm{N}_{[0.2032, 0.2032, 0.88, 21.84, 8.37, 21.84]}(\mu = 0.1 \times \mathrm{Dose}_t + 0.1 \times \mathrm{MEMS}_t + \mathrm{I}(\mathrm{Genotype}_0 \le 2) \times 2.66$$
$$+ \mathrm{I}(\mathrm{Genotype}_0 = 3) \times 4.6, \sigma = 4.06)$$

$$\mathrm{VL}_t \sim \mathrm{B}(p = 1 - (1/(1 + \exp(-(1 - 0.6 \times \mathrm{I}(t = 1) - 1.2 \times \mathrm{I}(t = 4) + 0.1 \times \mathrm{CoMo}_{t-1}$$
$$+ (2 - 0.2 \times \mathrm{I}(t = 3)) \times \sqrt{\mathrm{EFV}_t})))))$$

$$\mathrm{MV}_t \sim \mathrm{B}(p = 1/(1 + \exp(-(-2.95 + 0.1 \times \mathrm{SES}_0 + \mathrm{MV}_{t-1}))))$$

$$\mathrm{M}_{\mathrm{MEMS}_t} \sim \mathrm{B}(p = 1/(1 + \exp(-(0.5 \times I(\mathrm{TI}_t = 1) + 0.2))))$$

For $t \ge 2$:

$$\mathrm{Dose}_t \sim \mathrm{MN} \begin{pmatrix} p1 = (1/(1 + \exp(-(4 + \mathrm{Dose}_{t-1} \times 0.5 + \sqrt{\mathrm{Weight}_t} \times 4 - \mathrm{Age}_0 \times 10))), \\ p2 = (1/(1 + \exp(-(8 + \mathrm{Dose}_{t-1} \times 0.5 + \sqrt{\mathrm{Weight}_t} \times 8.568 - \mathrm{Age}_0 \times 9.06))), \\ p3 = (1/(1 + \exp(-(20 + \mathrm{Dose}_{t-1} \times 0.5 + \sqrt{\mathrm{Weight}_t} \times 6.562 - \mathrm{Age}_0 \times 18.325))), \\ p4 = 1 - (p1 + p2 + p3) \end{pmatrix}$$

The DGP for $\mathbf{G}_{alt1}$ (including the blue dashed lines in Fig. 3) coincides with the DGP above, except for the structural equations for MV_t, $t \in \{0, 6, 36, 48, 60, 84\}$. These are specified as follows for $\mathbf{G}_{alt1}$:

For $t = 0$:

$$\mathrm{MV}_0 \sim \mathrm{B}(p = 1/(1 + \exp(-(-2.95 + 0.1 \times \mathrm{SES}_0 + 2 \times \mathrm{VL}_0))))$$

For $t \ge 1$:

$$\mathrm{MV}_t \sim \mathrm{B}(p = 1/(1 + \exp(-(-2.95 + 0.1 \times \mathrm{SES}_0 + \mathrm{MV}_{t-1} + 2 \times \mathrm{VL}_t))))$$

Thus, MV_t additionally depends on VL_t, $t \in \{0, 6, 36, 48, 60, 84\}$, which corresponds to the dashed blue lines in the DAG.

The DGP for $\mathbf{G}_{alt2}$ (where SES is a cause of $MEMS_t$, $t \in \{6, 36, 48, 60, 84\}$) coincides with the DGP for $\mathbf{G}_{main}$ above, except for the structural equations for $MEMS_t$, $t \in \{6, 36, 48, 60, 84\}$. These are specified for $\mathbf{G}_{alt2}$ as follows:
For $t \ge 1$:

$$\mathrm{MEMS}_t \sim \mathrm{B}(p = 1/(1 + \exp(-(0.71 + \mathrm{CoMo}_{t-1} \times 0.31 + \mathrm{MEMS}_{t-1}$$
$$\times \mathrm{I}(t \ge 2) \times 0.31 - \mathrm{SES}_0 \times 0.5))))$$

After generating the data set using the structural equations, we introduce missing values based on the missingness indicators: if a missingness indicator equals 1, the corresponding covariate value is set to NA. In *Simulation 1*, we ignore the missingness indicators for EFV_t, $t \in \{6, 36, 48, 60, 84\}$, and $Weight_t$, $t \in \{0, 6, 36, 48, 60, 84\}$, and generate missingness only in VL_t and $MEMS_t$, $t \in \{0, 6, 36, 48, 60, 84\}$. In *Simulation 2*, missingness is introduced in EFV_t, $t \in \{6, 36, 48, 60, 84\}$, $Weight_t$, VL_t and $MEMS_t$, $t \in \{0, 6, 36, 48, 60, 84\}$.

D.2.   Results

The results below are based on Simulation 2 as defined in Section 5.

The following results are based on simulation 1 as defined in Section 5 under the assumption of $\mathbf{G}_{alt2}$ being the true underlying causal m-DAG.
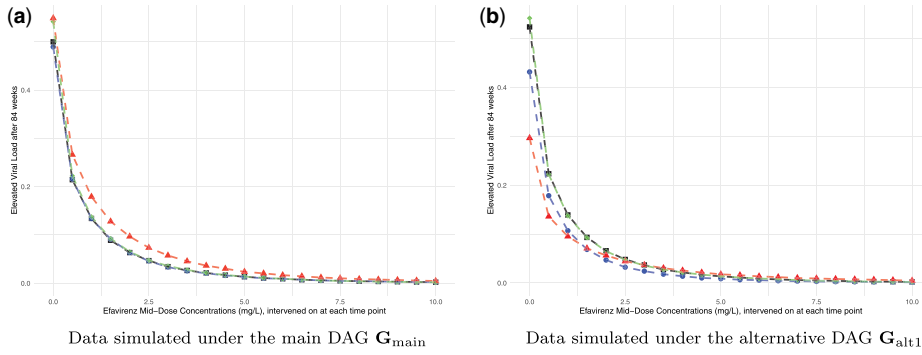
Figure D.1: Estimated CCRCs for the probability of viral failure after 84 wks. (a) Data simulated under the main DAG $\mathbf{G}_{main}$. (b) Data simulated under the alternative DAG $\mathbf{G}_{alt1}$. Causal effects were estimated on complete data (black squares), incomplete data using available cases (blue dots), incomplete data using multiple imputation (red triangles) and counterfactual data (green diamonds, true CCRC); results represent the mean over 1000 seeds.
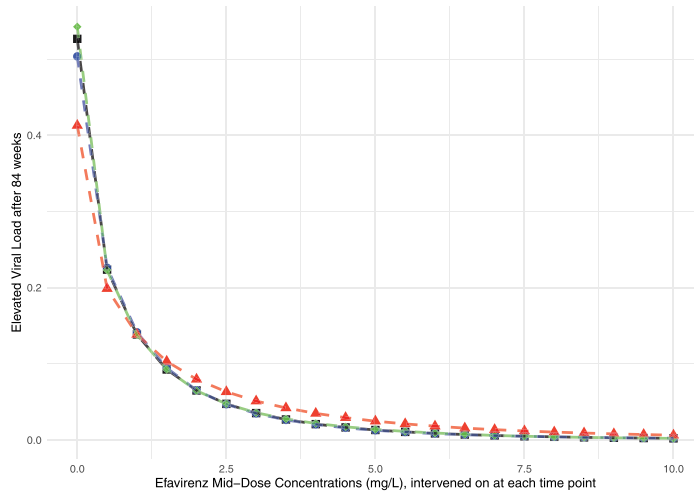

Figure D.2: Estimated CCRCs for the probability of viral failure after 84 wks based on data simulated under the main DAG $\mathbf{G}_{alt2}$. Causal effects were estimated on complete data (black squares), incomplete data using available cases (blue dots), incomplete data using multiple imputation (red triangles) and counterfactual data (green diamonds, true CCRC); results represent the mean over 1000 seeds.

D.3. Monte Carlo confidence intervals

The Monte Carlo confidence intervals in Tables D.1a, D.1b, and D.1c are reported to determine whether the differences between MI estimates and true values (main and alternative m-DAGs, $\mathbf{G}_{main}$ and $\mathbf{G}_{alt1}$), and between available case analysis estimates and true causal effects θ_{84} (alternative m-DAG $\mathbf{G}_{alt1}$) stem from simulation uncertainty or indicate systematic deviations. A 95% confidence interval [mean difference $-2 \cdot SE$; mean difference $+2 \cdot SE$], with the standard error SE computed as $sd(estimate)/\sqrt{(\#runs)}$, excluding zero suggests that the differences can likely not be explained by simulation uncertainty, indicating a bias caused by the estimation approach.

Table D.1. Monte Carlo confidence intervals for the difference between estimates and true causal effects across efavirenz (EFV) concentrations. (a) Multiple imputation estimates under the main m-DAG $\mathbf{G}_{main}$. (b) Available case estimates under the alternative m-DAG $\mathbf{G}_{alt1}$. (c) Multiple imputation (MI) estimates under the alternative m-DAG $\mathbf{G}_{alt1}$.

(a) Multiple imputation under $\mathbf{G}_{main}$			(b) Available case under $\mathbf{G}_{alt1}$			(c) Multiple imputation under $\mathbf{G}_{alt1}$		
EFV	Lower	Upper	EFV	Lower	Upper	EFV	Lower	Upper
0.0	-0.19441	-0.19351	0.0	−0.12787	−0.10112	0.0	−0.29980	−0.29904
0.5	-0.04522	-0.04452	0.5	−0.04754	−0.03497	0.5	−0.10118	−0.10062
1.0	-0.00881	-0.00824	1.0	−0.03208	−0.02512	1.0	−0.04923	−0.04876
1.5	0.00696	0.00749	1.5	−0.02465	−0.02075	1.5	−0.02354	−0.02312
2.0	0.01493	0.01539	2.0	−0.01873	−0.01642	2.0	−0.00888	−0.00849
2.5	0.01774	0.01817	2.5	−0.01470	−0.01320	2.5	−0.00111	−0.00076
3.0	0.01842	0.01880	3.0	−0.01195	−0.01092	3.0	0.00326	0.00358
3.5	0.01812	0.01847	3.5	−0.01012	−0.00935	3.5	0.00583	0.00614
4.0	0.01721	0.01754	4.0	−0.00823	−0.00759	4.0	0.00727	0.00755
4.5	0.01626	0.01656	4.5	−0.00677	−0.00620	4.5	0.00804	0.00830
5.0	0.01502	0.01530	5.0	−0.00558	−0.00509	5.0	0.00832	0.00857
5.5	0.01403	0.01429	5.5	−0.00450	−0.00405	5.5	0.00849	0.00871
6.0	0.01309	0.01335	6.0	−0.00358	−0.00317	6.0	0.00851	0.00872
6.5	0.01191	0.01214	6.5	−0.00296	−0.00257	6.5	0.00815	0.00835
7.0	0.01123	0.01145	7.0	−0.00222	−0.00186	7.0	0.00819	0.00838
7.5	0.01051	0.01071	7.5	−0.00175	−0.00141	7.5	0.00795	0.00814
8.0	0.00952	0.00971	8.0	−0.00145	−0.00115	8.0	0.00749	0.00766
8.5	0.00886	0.00904	8.5	−0.00114	−0.00085	8.5	0.00718	0.00734
9.0	0.00810	0.00827	9.0	−0.00088	−0.00062	9.0	0.00675	0.00691
9.5	0.00750	0.00766	9.5	−0.00074	−0.00048	9.5	0.00640	0.00655
10.0	0.00685	0.00701	10.0	−0.00058	−0.00034	10.0	0.00605	0.00620

AUTHOR CONTRIBUTIONS

The study was designed by MS, PD, and HM. Data was collected and interpreted by PD, and HM. All authors reviewed the study design and interpreted the data. Methods development was lead by AH. The first draft of the article was written by AH. All authors reviewed this and the following drafts, revising it critically for its content. All authors have read and approved the final version and agree with the manuscript's conclusion.

SUPPLEMENTARY MATERIAL

Supplementary material is available at Biostatistics Journal online.

CONFLICT OF INTEREST

None declared.

DATA AVAILABILITY

The code used for conducting the simulation study in Section 5 is accessible at the following GitHub repository: https://github.com/aholovchak/Recoverability-of-causal-effects.

REFERENCES

Abongomera G, Cook A, Musiime V, Chabala C, Lamorde M, Abach J, Thomason M, Mulenga V, Kekitiinwa A, Colebunders R, et al. 2017. Improved adherence to antiretroviral therapy observed among HIV-infected children whose caregivers had positive beliefs in medicine in Sub-Saharan Africa. Aids Behav. 21:441–449.

Balzer LB, van der Laan M, Ayieko J, Kamya M, Chamie G, Schwab J, Havlir DV, Petersen ML. 2021. Two-stage TMLE to reduce bias and improve efficiency in cluster randomized trials. Biostatistics. 24:502–517. URL https://doi.org/10.1093/biostatistics/kxab043.

Bienczak A, Denti P, Cook A, Wiesner L, Mulenga V, Kityo C, Kekitiinwa A, Gibb DM, Burger D, Walker AS, et al. 2016. Plasma efavirenz exposure, sex, and age predict virological response in HIV-infected African children. J Acquir Immune Defic Syndr. 73:161–168.

Hernán M, Robins J. 2020a. Causal inference: what if. Boca Raton: Chapman & Hall/CRC.

Hernan M, Robins J. 2020b. Causal inference. Boca Raton: Chapman & Hall/CRC. https://www.hsph.harvard.edu/miguel-hernan/causal-inference-book/

Hernán M, Robins J. 2006. Estimating causal effects from epidemiologic data. J Epidemiol Community Health. 60:578–586. https://doi.org/10.1136/jech.2004.029496.

Honaker J, King G, Blackwell M. 2011. Amelia II: a program for missing data. J Stat Softw. 45:1–47.

Mohan K, Pearl J. 2014. Graphical models for recovering probabilistic and causal queries from missing data. In: Ghahramani Z, Welling M, Cortes C, Lawrence N, Weinberger K, editors. Advances in neural information processing systems. Vol. 27. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2014/file/31839b036f63806cba3f47b93af8ccb5-Paper.pdf.

Mohan K, Pearl J. 2021. Graphical models for processing missing data. J Am Stat Assoc. 116:1023–1037. https://doi.org/10.1080/01621459.2021.1874961.

Mohan K, Pearl J, Tian J. 2013. Graphical models for inference with missing data. In: Burges C, Bottou L, Welling M, Ghahramani Z, Weinberger K, editors. Advances in neural information processing systems. Vol. 26. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2013/file/0ff8033cf9437c213ee13937b1c4c455-Paper.pdf

Moreno-Betancur M, Lee KJ, Leacy FP, White IR, Simpson JA, Carlin JB. 2018. Canonical causal diagrams to guide the treatment of missing data in epidemiologic studies. Am J Epidemiol. 187:2705–2715. https://doi.org/10.1093/aje/kwy173.

Mulenga V, Musiime V, Kekitiinwa A, Cook AD, Abongomera G, Kenny J, Chabala C, Mirembe G, Asiimwe A, Owen-Powell E, et al. 2016. Abacavir, zidovudine, or stavudine as paediatric tablets for African HIV-infected children (chapas-3): an open-label, parallel-group, randomised controlled trial. Lancet Infect Dis. 16:169–79.

Nugent JR, Marquez C, Charlebois ED, Abbott R, Balzer LB. 2024. Blurring cluster randomized trials and observational studies: Two-Stage TMLE for subsampling, missingness, and few independent units. Biostatistics. 25:599–616. https://doi.org/10.1093/biostatistics/kxad015. Erratum in: Biostatistics. 2024 Aug 26:kxae029. https://doi.org/10.1093/biostatistics/kxae029.

Pearl J. 2009. Causality 2nd ed. Cambridge, United Kingdom: Cambridge University Press. https://doi.org/10.1017/CBO9780511803161.

Pearl J, Verma TS. 1990. Equivalence and synthesis of causal models. In: Proceedings of the 6th Conference on Uncertainty in Artificial Intelligence. Elsevier. p. 220–227.

Robins J. 1986. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. Math Modell. 7:1393–1512. https://doi.org/10.1016/0270-0255(86)90088-6.

Rubin DB. 1976. Inference and missing data. Biometrika. 63:581–592.

Schafer JL, Graham JW. 2002. Missing data: our view of the state of the art. Psychol Methods. 7:147–177.

Schomaker M. 2020. Regression and causality, arXiv, https://arxiv.org/abs/2006.11754, preprint: not peer reviewed.

Schomaker M, Hogger S, Johnson LF, Hoffmann CJ, Barnighausen T, Heumann C. 2015. Simultaneous treatment of missing data and measurement error in hiv research using multiple overimputation. Epidemiology. 26:628–636.

Schomaker M, McIlleron H, Denti P, Dìaz I. 2023. Causal inference with continuous multiple time point interventions. Statistics in Medicine. https://doi.org/10.1002/sim.10246

Schomaker M, Denti P, Bienczak A, Burger D, Diaz I, Gibb D, Walker S, McIlleron H. 2024. Determining Targets for Antiretroviral Drug Concentrations: A Causal Framework Illustrated with Pediatric Efavirenz Data from the Chapas-3 Trial. Pharmacoepidemiology and Drug Safety. In press.

Seaman S, Galati J, Jackson D, Carlin J. 2013. What is meant by "missing at random"? Stat Sci. 28:257–268. https://doi.org/10.1214/13-sts415.

Shpitser I, Pearl J. 2006. Identification of joint interventional distributions in recursive semi-markovian causal models. In: Proceedings of the 21st National Conference on Artificial Intelligence, Vol. 2, AAAI'06. AAAI Press. p. 1219–1226.

Sofrygin O, van der Laan MJ, Neugebauer R. 2017. simcausal R package: conducting transparent and reproducible simulation studies of causal effect estimation with complex longitudinal data. J Stat Softw. 81:1–47. https://doi.org/10.18637/jss.v081.i02.

Tian J. 2015. Missing at random in graphical models. In: Lebanon G, Vishwanathan SVN, editors. Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics, Vol. 38 of Proceedings of Machine Learning Research, San Diego, California, USA, 09–12 May 2015. PMLR, p. 977–985. https://proceedings.mlr.press/v38/tian15.html

Tian J, Pearl J. 2002. A general identification condition for causal effects. In: Eighteenth National Conference on Artificial Intelligence, USA. American Association for Artificial Intelligence, p. 567–573.

Tikka S, Karvanen J. 2017. Identifying causal effects with the R package causaleffect. J Stat Softw. 76:1–30. https://doi.org/10.18637/jss.v076.i12.

van der Laan M, Rose S. 2011. Targeted learning: causal inference for observational and experimental data. Springer series in statistics. New York: Springer. https://books.google.de/books?id=RGnSX5aCAgQC