Missing Data 3.0: Can Missingness-DAGs Inform Statistical Analysis Choices in Longitudinal HIV Treatment Studies?

Anastasiia Holovchak, Paolo Denti, Elizabeth Kauda, Di Gibb, Sarah Walker, Helen McIlleron, Michael Schomaker

Background

Missing data are common in HIV treatment research. Both complete case analyses (missing data 1.0) and multiple imputation (missing data 2.0) are popular ways to address missing data, at least if certain missing (completely) at random assumptions are met. However, arguing for (or against) those assumptions in complex longitudinal settings is difficult – and which approaches are valid if the missingness mechanism is not at random, if any, remains often unclear. Recently, missingness directed acyclic graphs (m-DAGs) have been propsed to address these gaps.

Aims

We investigate the applicability of the framework of graphical models for handling missing data to a complex longitudinal pharmacoepidemiologic study of HIV-positive children treated with an efavirenz-based regimen (EFV) as part of the CHAPAS-3 trial, which enrolled children <13 years in Zambia/Uganda.

We ask: Can m-DAGs make their way from blackboards to actual applications?

Data

Our data comes from the CHAPAS-3 trial, which enrolled 478 HIV-positive children, under 13 years of age, in 4 sites in Uganda and Zambia. Children in the study received cART comprising two nucleoside reverse transcriptase inhibitors (lamivudine and randomly assigned abacavir, or stavudine or zidovudine) and one non-nucleoside reverse transcriptase inhibitor (efavirenz [EFV] or nevirapine). *We evaluated data of 125 children who received efavirenz – at 6, 36, 48, 60 and 84 weeks of follow-up.*

Scientific Question and Estimand

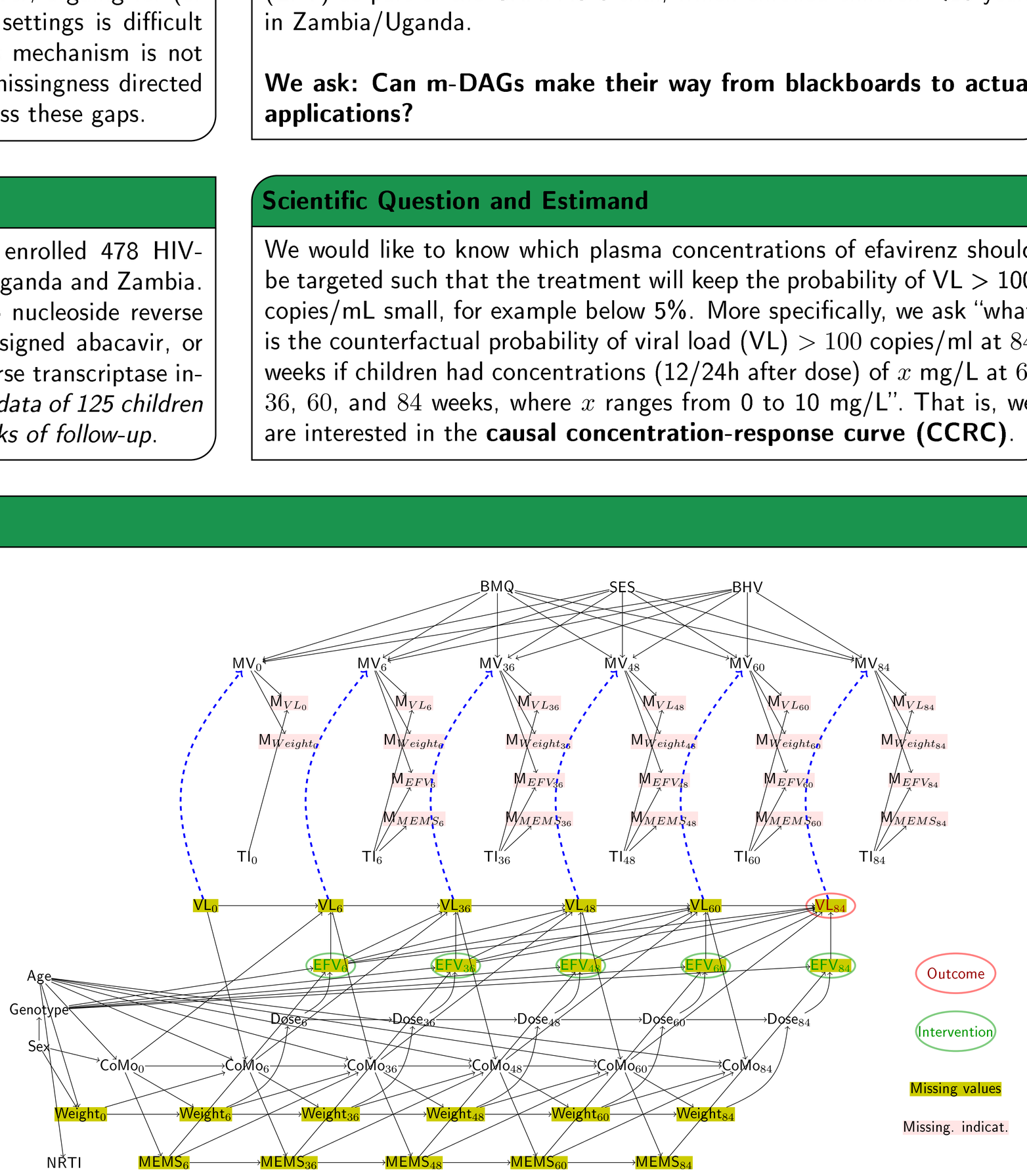
We would like to know which plasma concentrations of efavirenz should be targeted such that the treatment will keep the probability of VL > 100 copies/mL small, for example below 5%. More specifically, we ask "what is the counterfactual probability of viral load (VL) > 100 copies/ml at 84 weeks if children had concentrations (12/24h after dose) of x mg/L at 6, 36, 60, and 84 weeks, where x ranges from 0 to 10 mg/L". That is, we are interested in the **causal concentration-response curve (CCRC)**.

The Causal Missingness Model

We summarized clinician's knowledge on why data are possibly missing in a m-DAG (Figure).
The causal missingness graph contains

1. variables which are important to identify the effect of interest (i.e. the effect of EFV on viral failure (VL), bottom). $\rightarrow c-$DAG
2. binary missingness indicator variables (top, pink shading) for relevant variables with missing data (EFV, VL, adherence (MEMS), weight). $\rightarrow m-$DAG.

Reasons for missing data are represented by arrows leading to those indicators and include technical issues (TI, unmeasured) and missed visits (MV), which are themselves related to socio-economic status (SES, measured), beliefs and attitudes towards medicine (BMQ, measured) and other behavioural factors (unmeasured). Additional speculative reasons for missed visits are represented by blue dashed arrows.



General Results

- Under the assumptions encoded in the m-DAG, 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. Also, unmeasured behavioral factors may cause missed visits.
- However, we show that **the assumptions are sufficient to estimate the CCRC by means of the available data, using specific g-formula type-of representations – despite MNAR**.
- If missed visits would be caused by the outcome (elevated viral load, dashed blue arrows), the causal effect can however not be recovered (no indication from clinicians though).
- Interestingly, additional simulations show that recoverability holds true even if behavioural factors directly cause adherence patterns.
- Estimated concentration-response curves (Figure, right) are much flatter after multiple imputation – and are actually invalid, as predicted by theory and confirmed by us in simulations.

Analysis Results

RED = Imputation

BLUE = Available Cases

Conclusions

- We have shown the applicability of m-DAGs to complex longitudinal studies.
- Our analyses and derivations demonstrate that sometimes an available case analyses can be valid under MNAR, while imputation is invalid.
- However, our application also highlights the massive effort involved, technical expertise required and sensitivity of results with respect to the assumed causal model.

Affiliations & Preprint

Affiliations: AH: Seminar for Statistics, ETH Zürich, Switzerland; MS: Department of Statistics, University of Munich (LMU), Germany; PD & HMI: Division of Clinical Pharmacology, University of Cape Town, South Africa; EK: Joint Clinical Research Centre, Uganda; DG & SW: MRC Clinical Trials Unit, University College London, United Kingdom .

Contact:
michael.schomaker@stat.uni-muenchen.de

Read our preprint: