Time for a change: considering regimen changes in analyses of observational MDR/RR-TB treatment cohort data

Running title: Regimen changes in MDR-TB treatment analyses

Molly F. Franke,[1] Carole D. Mitnick[1,2]

[1]Department of Global Health Social Medicine, Harvard Medical School, Boston, MA 02118

[2]Division of Global Health Equity, Brigham and Women's Hospital, Boston, MA 02118

Corresponding author:

Molly F. Franke

Department of Global Health Social Medicine

Harvard Medical School

641 Huntington Avenue

Boston, MA 02118

molly_franke@hms.harvard.edu

(617) 432-5224

Abstract: Randomized clinical trials represent the gold standard in therapeutic research. Nevertheless, observational cohorts of patients treated for multidrug-resistant (MDR) or rifampin-resistant (RR) tuberculosis (TB) also play an important role in generating evidence to guide MDR/RR TB care. Generally, summary exposure classifications (e.g., 'ever versus never', 'exposed at baseline') have been used to characterize drug exposure, in the absence of detailed longitudinal data on MDR-TB regimen changes. These summary classifications, along with an absence of data on covariates that change throughout the course of treatment, constrain researchers' ability to answer the most relevant questions while accounting for known biases. This paper highlights the importance of regimen changes in improving inference from observational studies of longer MDR-TB treatment regimens and offers an overview of the data and analytic strategies required to do so.

Keywords: tuberculosis, epidemiology, drug-treatment, resistance, analysis

Randomized controlled trials (RCT) of treatments for multidrug-resistant (MDR) and rifampicin-resistant (RR) tuberculosis (TB) play a critical role in generating high quality data on which to base patient care. RCTs represent the gold standard in this research for good reasons, including an ability to generate similar groups for comparison, thereby minimizing confounding due to measured and unmeasured variables. Recent efforts to improve the timeliness and relevance of RCTs will further enhance the value of these data.[1,2] Observational studies of patients treated under routine programmatic conditions are also critically important to respond to urgent questions for which RCT results are not yet available. Observational research can also examine whether findings from RCTs, with their often rigorous inclusion criteria, generalize to the majority of patients with TB, including patients from high-risk, often-underrepresented subgroups. Observational datasets may also shed light on questions for which there are no planned RCTs, such as optimal drug substitutions and management strategies in the presence of adverse events. For decades, observational data have been the primary evidence base informing MDR-TB treatment guidelines,[3,4] and this is likely to be true for the foreseeable future. Therefore, the ability to generate robust evidence from observational studies, as well as RCTs, will ensure that patient care is based on high quality evidence regardless of the study design. This paper highlights the importance of regimen changes in improving inference from observational studies of longer MDR-TB treatment regimens and offers an overview of the data and analytic strategies required to do so.

Longer (18-20 month) MDR-TB regimens are dynamic, with drug changes occurring frequently throughout the course of treatment. In the endTB observational study, a prospective cohort of patients receiving bedaquiline (BDQ) and/or delamanid (DLM) as part of a longer regimen for RR/MDR-TB in one of 17 countries,[5] patients received a median of five [interquartile range, IQR: 3 to 7] unique drug combinations during the course of treatment, and had a median of 8 [IQR: 5 to 11] regimen changes, including dose adjustments and temporary suspensions.[11] While some modifications to longer MDR-TB regimens are planned (e.g., BDQ dose adjustment after two weeks, injectable agents stopped after the intensive phase; BDQ or DLM stopped after 24 weeks), patient clinical evolution also drives changes (e.g., when sputum has not yet converted from positive to negative; due an adverse event). Figure 1 shows the number of drugs

[1] Calculated from the subset of patients who initiated a treatment with endTB prior to April 1, 2017.

received throughout the course of treatment in a random sample of 39 patients enrolled in the endTB observational study. Changes occurred throughout treatment, even beyond the eight-month intensive phase after which regimen changes might be expected to be rare: only 44% percent of patients received the same drugs at the end of treatment that they had received at 9 months. And, in 15% of patients at least two drugs changed between month 9 and the end of treatment. While drug subtractions were the most common changes, some patients also had drugs added to their regimens.

A lack of detailed regimen data from MDR-TB treatment cohorts has constrained the questions that can be answered and the ability to control for known biases. Historically, researchers have lacked standardized longitudinal regimen data from observational cohorts. Many MDR-TB treatment datasets consist of routinely collected programmatic data with limited data points. The individual patient database, perhaps the largest and richest source of MDR-TB treatment data, consists of pooled retrospective and prospective data contributed by TB programs and research studies from all over the world.[6] A logistical implication of merging heterogeneous data types is that patient data from studies with highly-detailed data (e.g., regimen details for each day of treatment) must be condensed to the lowest level of granularity in the pooled data (e.g., whether the patient was ever or never exposed to a drug) in order to conduct analyses. The resulting pooled large datasets offer the advantage of increased statistical power and patient diversity, but at the cost of a loss of detail that may influence or worse, determine, how and why regimens may change over time.

Figure 2 shows illustrative data for six patients who received a drug of interest, DLM for example, for varying durations (i.e., patient A received the drug for the entire duration of treatment; patient B received the drug for one month, patient C received the drug for six months, and patients D-F had the drug added to their regimen later on in treatment). An 'ever versus never' analysis of DLM exposure cannot distinguish between the duration and timing of exposure, which is highly variable across these patients. Thus, a key disadvantage of the 'ever versus never' characterization is the interpretation of the effect estimate (e.g., risk ratio, odds ratio), which will represent a mix of these highly variable DLM exposures. For this effect estimate to be meaningful, either all exposed patients must be similarly exposed to DLM (this is not the case for these six patients) *or* the duration and timing of DLM must have no bearing on effectiveness (an

implausible assumption in a study of effectiveness). Furthermore, variability in patterns of use between cohorts will challenge comparability of findings across studies that have ostensibly used the same summary exposure classification. A second limitation to the 'ever versus never' classification is immortal person-time bias, a bias relevant to cohorts in which patients commonly have a drug added to the regimen after initiation of follow-up (i.e., post-baseline).[7] Because patients who initiate drugs later in treatment must survive to the time at which they initiate the drug, an 'ever versus never' classification may overestimate a beneficial drug effect, suggest a beneficial effect when one does not exist, or underestimate a harmful effect.

An alternative to the 'ever versus never' classification is characterization of the regimen based on the drugs received at the beginning of treatment (i.e., "baseline" or "time 0"). While all of the patients in Figure 2 received the drug of interest, only patients A - C received the drug from the beginning of follow-up. In an analysis that characterizes regimens according to the baseline regimen, patients D - F would be classified as unexposed to the drug. As with the 'ever versus never' classification, the relevancy of this exposure summary measure to the causal effect of interest is questionable given that patients were treated for varying durations (e.g. 1 month, 6 months, or the entire duration of treatment). Furthermore, classifying patients based on baseline exposure fails to acknowledge DLM exposure in patients that had the drug added later on. This makes the two comparison groups more similar in terms of their exposure to DLM, likely attenuating any effect of DLM and reducing statistical power.

The challenges of the 'ever versus never' and 'baseline' classifications are not limited to studying treatment effects. These classification strategies are also problematic for any drug that is a potential confounder and requires adjustment in analyses. For example, if an investigator has data on baseline drug exposure only, and DLM was frequently added to regimens later on (e.g., patients C-F in Figure 2), an absence of data on post-baseline DLM use would preclude complete adjustment for DLM, resulting in residual confounding. Under the 'ever versus never' model, one cannot discern when DLM is a potential confounder (i.e., received along with the drug of interest) and when DLM was received after the drug of interest (i.e., substitution). This distinction is important. In the former scenario, adjustment is necessary and, in some circumstances, use of conventional regression techniques is a valid approach. In the latter, adjustment affects the

interpretation of the effect estimate and conventional regression techniques may induce bias.[8]

How likely are biases due to summary exposure classifications? The potential for summary measure exposure classifications to induce bias depends on the extent to which drug changes occur throughout treatment. Let's continue with the example of DLM. In the endTB observational study, DLM administration varied throughout treatment for at least three reasons. First, the availability of DLM to country programs increased over time. Second, DLM was added to or substituted into regimens as a result of evolving clinical events, such as new DST results or adverse events. Third, as physicians became more comfortable with the drug, and more comfortable using it with BDQ, they were more likely to prescribe it. Among patients receiving treatment in the endTB observational study, we classified DLM exposure according to commonly-used summary exposure measures and found that 1,117 received DLM on day one, 1,144 ever received DLM in first month, 1,363 had received DLM for at least a month, and 1,407 had ever received DLM.[22] Importantly, there was a 21% difference in number of people that met the most common (ever received DLM) and least common (received DLM on day one) exposure classifications. At least some of the reasons for regimen changes in the endTB observational cohort (i.e., suboptimal clinical evolution, side effects, changes in drug availability) are widely generalizable to clinical settings and thus it likely that there is at least some bias from summary exposure measures in most existing MDR-TB treatment studies that use them. The magnitude of this bias and any impact on overall study conclusions is impossible to know.

What is needed for the 'ideal' analyses of longer MDR-TB treatment regimens? Detailed longitudinal data on treatment changes, specifically drug start and stop dates, obviate the need for a single summary measure to classify a patient's drug exposure. When these data are available, each patient's drug exposure can be classified to reflect the exposure they actually received. In addition to detailed data on drug exposure and baseline confounders, data are needed on the post-baseline factors that predict drug exposure and outcome. Also known as potential time-varying confounders, these predictors include indicators of clinical evolution (e,g, new culture results or DST results), other drugs in the regimen, and adverse events. Conventional methods may not be appropriate for adjusting for time-varying confounders;[9] however adequate methods

[2] Calculated from among all observational study participants, including some of whom were still on treatment.

(i.e., inverse probability weighting, g-formula, g estimation of a structural nested model[10–12]), and guidance and resources for implementing them (e.g., code for programming these analyses in statistical software packages such as SAS, Stata and R), are becoming mainstream,[13] facilitating their use.

A cautionary example from HIV. Studies from other disciplines, including HIV, have found that observational studies may be biased when investigators disregard time-varying confounders. Hernán and colleagues offer an illustrative example in analyses of observational data on of the effect of zidovudine (ZDV) on mortality among men living with HIV in the late 1990s, a time when ZDV was indicated for patients with advanced HIV.[14] In that study men initiated ZDV throughout follow-up, and investigators had detailed longitudinal data on ZDV start date and the factors that determined ZDV initiation (i.e., CD4 cell count and viral load). In unadjusted analyses, investigators found that ZDV was positively associated with death, indicating a harmful effect. This is a classic example of confounding by indication: ZDV was preferentially administered to the sickest patients. In a second analysis adjusting for baseline confounders, the positive association was attenuated, but ZDV remained strongly associated with mortality. It was only when the investigators adjusted for the "on treatment" characteristics that predict ZDV use (in other words, the fact that the patients who started ZDV had lower CD4 cell counts and higher HIV diseases stages) that they saw the protective association between ZDV and mortality that had been observed in clinical trials.

Time for a change in RR/MDR-TB. A recent review found that, since 2000, there have been over 400 papers have used the above methods to appropriately adjust for time-dependent confounding, with sharp inclines since 2013 and 112 articles published in 2016.[15] The authors reported that publications using these methods span many disciplines; however, they have been concentrated largely in the fields of HIV,[14,16,17] cardiopulmonary health,[18,19] kidney disease,[20,21] and mental health / neurology.[22,23] A review by our group found that the only TB-related papers that have used these methods to account for time-varying confounding, were papers that were foremost related to HIV.[24] We found no evidence that these methods have been used in analyses of MDR-TB treatments; this may be attributable in large part to the lack of longitudinal data that could be used to perform such analyses.

Conclusions. A lack of detailed longitudinal data from MDR-TB treatment cohorts has prevented the

generation of high-quality data that enables research to respond to the most pressing questions related to MDR-TB treatment. The magnitude and direction of biases in existing analyses are difficult to predict, and it is impossible to know how the ensuing results may affect current patient care. Improving analyses of observational data will require large detailed longitudinal datasets, along with collaboration and coordination to harmonize and streamline data collection and data management and financial investment to support the programs and teams collecting these data. Some of these efforts are already underway.[25,26] Thorough statistical analyses, including sensitivity analyses, must be implemented by investigators, and reinforced by peer-reviewers and editorial boards. Finally, additional methods development will be needed to adapt existing methods to accommodate the complexities of MDR-TB. These steps in combination with existing tools, such as the target trial framework (i.e., the design of observational analyses that emulate the ideal clinical trial[18,19,27–30]) and directed acyclic graphs (i.e., causal diagrams that facilitate identification of potential confounders and other potential sources of bias[8,31,32]), will further hone questions and analyses, ensuring that patients with MDR-TB receive evidence-based care driven by high-quality evidence, even when RCT data is unavailable.

Acknowledgments

We acknowledge and thank the endTB observational study team, and especially Sid Atwood for his programming support. The authors declare no conflict of interest.

Conflict of Interest: The authors declare no conflict of interest.

References

1 Cellamare M, Ventz S, Baudin E, Mitnick CD, Trippa L. A Bayesian response-adaptive trial in tuberculosis: The endTB trial. *Clin Trials* 2017; **14**: 17–28.

2 Phillips PPJ, Mitnick CD, Neaton JD, Nahid P, Lienhardt C, Nunn AJ. Keeping phase iii tuberculosis trials relevant:Adapting to a rapidly changing landscape. *PLoS Med* 2019; **16**. DOI:10.1371/journal.pmed.1002767.

3 World Health Organization. WHO treatment guidelines for multidrug- and rifampicin-resistant tuberculosis, 2018 update. *WHO* 2019.

4 WHO treatment guidelines for drug-resistant tuberculosis, 2016 update (WHO/HTM/TB/2016.04) [Internet]. Geneva, World Health Organization. 2016. Available from: http://www.who.int/tb/areas-of-work/drug-resistanttb/treatment/resources/en/. .

5 Khan U, Huerga H, Khan A, *et al.* Treatment of MDR-TB with bedaquiline or delamanid containing regimens. *BMC Infect Dis* 2019; **19**.

6 Ahmad N, Ahuja SD, Akkerman OW, *et al.* Treatment correlates of successful outcomes in pulmonary multidrug-resistant tuberculosis: an individual patient data meta-analysis. *Lancet* 2018; **392**: 821–34.

7 Suissa S. Immortal time bias in observational studies of drug effects. *Pharmacoepidemiol Drug Saf* 2007; **16**: 241–9.

8 Cole SR, Hernán MA. Fallibility in estimating direct effects. *Int J Epidemiol* 2002; **31**: 163–5.

9 Robins JM, Hernán MA, Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology* 2000; **11**: 550–60.

10 Hernán MA, Lanoy E, Costagliola D, Robins JM. Comparison of Dynamic Treatment Regimes via Inverse Probability Weighting. *Clin PharmacolToxicol* 2006; **98**: 237–42.

11 Naimi AI, Cole SR, Kennedy EH. An introduction to g methods. *Int J Epidemiol* 2017; **46**: 756–62.

12 Cole SR, Hernán MA. Constructing Inverse Probability Weights for Marginal Structural Models. *Am J Epidemiol* 2008; **168**: 656–64.

13 Hernán MA, Robins JM. Causal Inference: What If. Boca Raton: Chapman & Hall/CRC, 2020. Available from: https://www.hsph.harvard.edu/miguel-hernan/causal-inference-book/

14 Hernán MA, Brumback B, Robins JM. Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. *Epidemiology* 2000; **11**: 561–70.

15 Clare PJ, Dobbins TA, Mattick RP. Causal models adjusting for time-varying confounding - A systematic review of the literature. *Int J Epidemiol* 2019; **48**: 254–65.

16 Westreich D, Cole SR, Young JG, *et al.* The parametric g-formula to estimate the effect of highly active antiretroviral therapy on incident AIDS or death. *Stat Med* 2012; **31**: 2000–9.

17 Patel K, Hernán MA, Williams PL, *et al.* Long-term effectiveness of highly active antiretroviral therapy on the survival of children and adolescents with HIV infection: a 10-year follow-up study. *Clin Infect Dis* 2008; **46**: 507–15.

18 Danaei G, García Rodríguez LA, Cantero OF, Logan RW, Hernán MA. Electronic medical records can be used to emulate target trials of sustained treatment strategies. *J Clin Epidemiol* 2018; **96**: 12–22.

19 Danaei G, Rodríguez LAG, Cantero OF, Logan R, Hernán MA. Observational data for comparative effectiveness research: An emulation of randomised trials of statins and primary prevention of coronary heart disease. *Stat Methods Med Res* 2013; **22**: 70–96.

20 Cotter D, Zhang Y, Thamer M, Kaufman J, Hernán MA. The effect of epoetin dose on hematocrit. *Kidney Int* 2008; **73**: 347–53.

21 Lertdumrongluk P, Streja E, Rhee CM, *et al.* Dose of hemodialysis and survival: A marginal structural model analysis. *Am J Nephrol* 2014; **39**: 383–91.

22 Zhong QY, Gelaye B, VanderWeele TJ, Sanchez SE, Williams MA. Causal model of the association of social support with antepartum depression: Amarginal structural modeling approach. *Am J Epidemiol* 2018; **187**: 1871–9.

23 Bentley R, Baker E, Simons K, Simpson JA, Blakely T. The impact of social housing on mental health: longitudinal analyses using marginal structural models and machine learning-generated weights. *Int J Epidemiol* 2018; **47**: 1414–22.

24 Rodriguez CA, Sy KTL, Mitnick CD, Franke MF. What are we missing in analyses of tuberculosis treatment cohorts: a review of methods to control for time-dependent confounding. In: World Conference on Lung Health. Hyderabad, India, 2019.

25 Campbell JR, Falzon D, Mirzayev F, *et al.* Improving Quality of Patient Data for Treatment of Multidrug- or Rifampin-Resistant Tuberculosis. *Emerg Infect Dis* 2020; **26**. DOI:10.3201/eid2603.190997.

26 The ShORRT Research Package. https://www.who.int/tdr/research/tb_hiv/en/ (accessed Feb 7, 2020).

27 Hernán MA, Sauer BC, Hernández-Díaz S, Platt R, Shrier I. Specifying a target trial prevents immortal time bias and other self-inflicted injuries in observational analyses. *J Clin Epidemiol* 2016; **79**: 70–5.

28 Hernán MA, Robins JM. Using Big Data to Emulate a Target Trial When a Randomized Trial Is Not Available. *Am J Epidemiol* 2016; **183**: 758–64.

29 García-Albéniz X, Hsu J, Hernán MA. The value of explicitly emulating a target trial when using real world evidence: an application to colorectal cancer screening. *Eur J Epidemiol* 2017; **32**: 495–

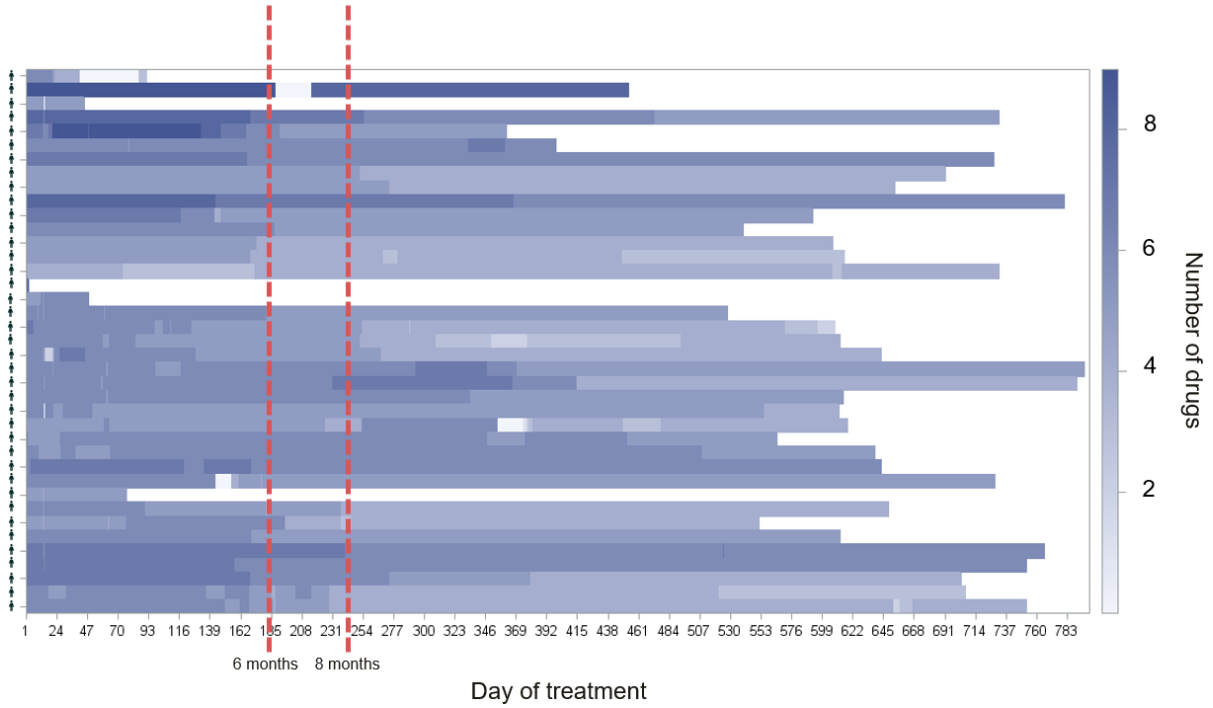
500.

30 Caniglia EC, Zash R, Jacobson DL, *et al.* Emulating a target trial of antiretroviral therapy regimens started before conception and risk of adverse birth outcomes. *AIDS* 2018; **32**: 113–20.

31 Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research. *Epidemiology* 1999; **10**: 37–48.

32 Hernán MA, Hernández-Díaz S, Robins JM. A structural approach to selection bias. *Epidemiology* 2004; **15**: 615–25.

234 Figure 1. Heatmap of the number of drugs received during the course of treatment for MDR/RR TB,
235 among 39 patients randomly selected from 13 countries in the endTB observational study.



236

237 Legend: Each horizontal bar represents a patient. The day of treatment is on the x-axis. The color legend
238 is on the right, with darker colors representing a higher number of drugs. Reference lines at six and eight
239 months indicate times at which prescribed changes to treatment often occur.

240

241

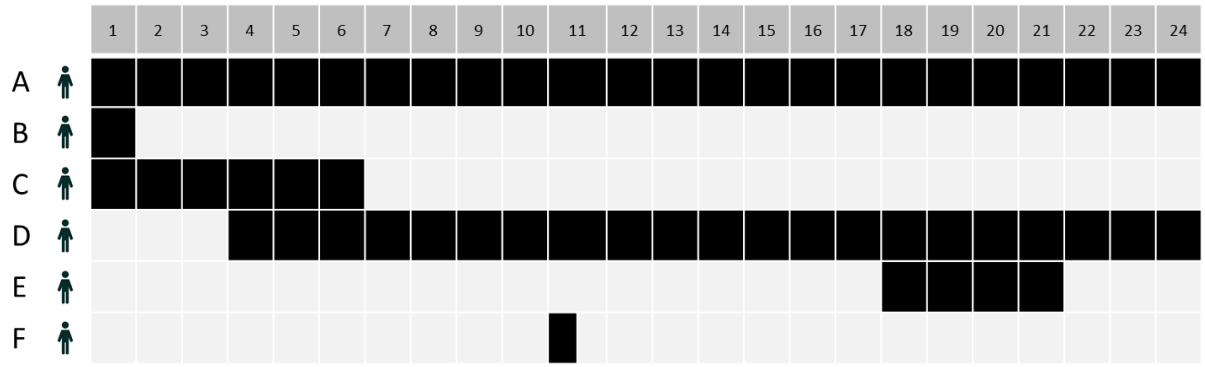
242

243

244

245

246 Figure 2. Illustrative data from six patients.



247

248 Legend: black shaded squares indicate months of drug exposure and light squares represent months ofno

249 drug exposure.