Assessing treatment effect heterogeneity using baseline risk

Methodology and applications

Alexandros Rekkas

2022-05-18

Contents

Preface **5**

1 Introduction **7**

Preface

Thesis

Chapter 1

Introduction

In order to provide—on average—the most current medical care doctors are advised to align their clinical practice with the results of well-conducted clinical trials, or the aggregation of the results from multiple such trials [Greenfield 2007]. This approach implicitly assumes that all patients eligible for treatment similarly experience the benefits and harms of treatment of the reference trial population. Therefore, at a certain point, the accurate estimation of these average effects became crucial, transforming clinical trials from tools for assessing causality into tools for predicting patient-level treatment effects. When the strong positive overall effects derived from clinical trials could not be achieved in medical practice, the problem was attributed to the reference trial population being too narrow and not representing the "average" patient requiring treatment. Therefore the need for more pragmatic clinical trials incorporating broad patient populations was highlighted [Treweek, Trials 2009; …].

The wider clinical trial populations ensure that overall results will be generalizable to the "real-life" patients. However, generalizability comes at a cost: wider range of included patients means higher variability of measured characteristics, therefore higher variability in disease severity is observed, which, in turn, translates to higher variability of observed treatment effect sizes. In short, the estimated treatment effect derived from such clinical trials is often an average of heterogeneous treatment effects and, as such, is not applicable to most patient subgroups. This means that a positive average treatment effect estimated from a clinical trial very often is only evidence that some of the enrollees benefitted from the intervention under study. If, however, the intervention is linked to a serious adverse event, treating everyone would result in serious harms for many patients, despite the positive overall effect.

Conceptually, heterogeneity of treatment effects (HTE) is the variation of treatment effects on the individual level across the the population [https://doi.org/10.1111/j.0887-378X.2004.00327.x]. As individual treatment effects—difference between outcomes under all possible treatment assignments within the same

individual—are the ones that generate HTE, its identification and quantification is crucial for guiding medical decision making and lies at the core of patient-centered research. Despite HTE being widely anticipated, however, its evaluation is not straightforward. Individual treatment effects are—by their nature—unobservable since, the moment a patient receives a specific treatment, their response under the alternatives becomes unmeasurable (fundamental problem of causal inference).

To "glimpse" at a specific individual's response under alternative treatments, researchers usually observe the outcomes of other "similar" patients that did receive one of the other candidate treatments. However, patient similarity is not straightforward to assess. Patients differ in a vast number of characterstics that make them unique and may or may not be relevant to modifying treatment responses. Identification of such patient characterstics can be quite complicated. In clinical trials it usually relies on the detection of measured covariates with a statistically significant interaction with treatment. However, as clinical trials are usually only adequately powered to detect overall effects of a certain size, this kind of analyses can be quite problematic. This is already widely known and a large part of research in subgroup analyses has provided guidance on how these should be carried out [refs]. This lack of statistical power often results in falsely concluding "consistency" of the treatment effect across several subpopulations of interest, or overestimating the effect size of a treatment-covariate interaction. The former because an existing interaction effect was smaller than the detectable effect size, the latter because randomization did not achieve balance between the levels of the subgrouping variable due to the small number of patients belonging to the specific subgroup.

they require knowledge of patient-level outcomes under all possible treatment assignments (fundamental problem of causal inference). For that reason an average effect derived from a patient subgroup is often used for making "individualized" evaluations. These subgoups are usually defined based on a single patient characteristic, comparing treatment effects between males and females, older and younger patients, and any other covariate assumed to be relevant. However, as clinical trials are most often adequately powered to detect a certain overall effect size, these subgroup analyses usually are underpowered. This can lead to falsely claiming absence of HTE or overestimating its magnitude [refs]. In addition, contrary to subgroup analyses, patients differ with regard to many baseline covariates simultaneously [Kent, BMJ 2018]. However, evaluation of two-way or higher order interactions becomes underpowered very fast.

Kravitz et al [https://doi.org/10.1111/j.0887-378X.2004.00327.x] linked prediction of individual treatment effect on the knowledge of the patient's baseline risk, responsiveness to treatment, vulnerability to side-effects, and the utilities the patient places for different outcomes. Therefore, baseline risk, i.e. the patient's true probability of having an outcome of interest—usually the outcome treatment is attempting to prevent— without receiving the treatment under study is already recognized as an important determinant of treatment effect. It

sets an upper bound on how large the treatment effect can be for individuals of the same risk. Low risk patients can only experience minimal treatment benefit before their risk is reduced to zero. On the other hand, high risk patients can achieve much higher benefits. But, depending on the scale treatment effect is measured, HTE may be present or absent. For example, despite finding statistically significant subgroup effect evaluated on the relative scale, the absolute risk difference between the two groups may be so small that has no clinical relevance [http://dx.doi.org/10.1016/j.jclinepi.2013.11.003]. Therefore, in the presence of a trully effective treatment, effect heterogeneity should always be anticipated on some scale [Dahabreh, IJE 2016], as baseline risk is bound to vary across trial patients.

In the presence of true effect modification, accurately modeling treatment-covariate interactions would result in adequate evaluation of HTE. However, simulations have shown that due to the low power for correctly estimating these interactions, these approaches require very large sample sizes. Modeling effect modifiers essentially suffers from the same issues as subgroup analyses.

Healthcare data routinely collected by general practitioners, hospitals, insurance companies, and many other private or public bodies is becoming increasingly available for research, giving researchers access to massive amounts of patient data. Theoretically, the aforementioned power issues for the evaluation of HTE would be largely mitigated if the analyses were performed on even a single observational database. However, as this data is not being accumulated with research in mind, it suffers from many biases causing most of the traditional inference methods to fail. The main source of all the problems is that observational data is not randomized. A doctor prescribing a specific treatment knows—usually based on results from RCTs—that it will be beneficial for the patient they are treating. This often causes patients of similar characteristics following the same treatment course, resulting in non-random differences between the several treatments. This makes comparisons of treatments quite difficult to evaluate.