Assessing treatment effect heterogeneity using baseline risk

Methodology and applications

Alexandros Rekkas

2022-05-20

Contents

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 as the reference trial population. Therefore, at a certain point, the accurate estimation of these average effects became crucial for guiding medical practice, transforming clinical trials from tools for assessing causality into tools for predicting patient-level treatment effects [ref]. 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. It is often the case that older patients or patients with comorbidities or receiving multiple medications are not well represented in many clinical trials. 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 "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, may not be 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 effect (HTE) is the variation

of treatment effects on the individual level within the target 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 generating HTE, its identification and quantification is crucial for guiding medical decision making and lies at the core of patient-centered outcomes research. Despite HTE being widely anticipated, however, its evaluation is not straightforward. Individual treatment effects are—by their nature—unobservable: 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 actually received 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 statistically significant interactions of treatment with measured covariates (subgroup analyses). However, as clinical trials are usually only adequately powered to detect overall effects of a certain size, this kind of analyses can be problematic. This is already widely known and research in subgroup analyses has provided guidance on how these should be carried out [refs]. The 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 patient numbers.

Baseline risk, i.e. the patient's true probability—given their measured characteristics—of experiencing the outcome of interest without receiving the treatment under study, is an important determinant of treatment effect [refs]. 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 contrary, high risk patients can achieve much higher benefits. This means that baseline risk can be used as a subgrouping variable for assessing HTE. For many common settings prediction models of high quality for estimating baseline risk already exist and can be directly applied to the data at hand [refs]. If no such models exist, the researcher can develop one from the available dataset [refs]. An important advantage of using baseline risk as the subgrouping variable is that it, being a summary score (combination of the effects of multiple patient characteristics), achieves higher power for detecting HTE as it divides the population in much denser subpopulations compared to multiple-way subgroup analyses based on the same patient characteristics.

Depending on the scale treatment effect is measured, HTE may or may not be

identified. 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. If true effect modifiers and their relationship with the outcome were known and the available sample size provided adequate statistical power for evaluating these interactions, then modeling treatment-covariate interactions would be the optimal approach for assessing HTE. However, this usually is not the case and prursuing this approach may lead to overfitting and unstable estimates for the interaction effect sizes. In this case the tradeoff of using a more stable but misspecified model (risk-based subgrouping approach) can provide a valid alternative.

Healthcare data routinely collected by general practitioners, hospitals, insurance companies, and many other private or public bodies is becoming increasingly available giving researchers access to massive amounts of patient data. Theoretically, the aforementioned statistical power challenges for the evaluation of HTE would be largely mitigated if the analyses were performed on a single observational database. However, as this data is not being accumulated for research purposes, it suffers from many biases causing traditional inference methods to fail. The issue is that observational data is not randomized. A doctor prescribing a specific treatment expects—usually based on results from clinical trials—that it will be beneficial for the patient they are treating. This causes systematic differences in important characteristics among patients receiving different treatments and renders their comparison very untrustworthy.

If it can be assumed that all relevant patient characteristics on which the treating physician based their decision have been captured in the observational dataset, there are methods that can be used to account for these systematic differences [refs]. Among the more popular ones is limiting the analyses to the propensity score matched subpopulation. Propensity scores are the patient probabilities of receiving the treatment under study and have been shown to have the balancing property, i.e. conditional on the propensity score treatment assignment is independent of the potential outcomes [refs]. This means that in a subset of patients with equal propensity scores there are no differences in covariate distributions of patients receiving the treatment under study compared to those who are not. Consequently, patients within this subset can be assumed to be randomized.

However, information captured in observational databases is fragmented depending on the interests of the one gathering the data. Different patient characteristics are captured—at different levels of detail—in a general practitioner database, in a hospital medical record or in an administrative claims database. Because of that, it cannot be certain that all the relevant information used by the practioner is available within the data at hand. As consequence, propensity score adjustment will not suffice to evaluate treatment effects using the observational data, be it overall or subgroup effects.

Another important problem with observational databases is that they are not compatible with each other. As anyone gathering routinely observed healthcare data did so in a way that was more convenient to them, a plethora of structures for the resulting databases arose. Diseases, treatments, medical exams and many more aspects of healthcare are often coded differently in different observational databases. This means that combining results from multiple databases is not a simple task.

One of the solutions put forward for handling database incompatibility was the creation of Observational Medical Outcomes Partnership Common Data Model (OMOP-CDM) [refs]. This provided a standard for structuring an observational database while large effort was put into developing processes for mapping existing databases from their own specific structure to OMOP-CDM. A new level of standardization and scalability for observational studies was reached. Common definitions of diseases, treatments, and outcomes can now be applied uniformly across a network of many databases containing information on hundreds of millions of patients. An analysis plan can be executed following the exact same steps across this network giving a more complete picture of the effect sizes that can be estimated in different populations. The fragmented information scattered across multiple databases can now be summarized in a consistent way to give a fuller picture.