Baseline risk in medical decision making

From outcome prediction to the assessment of treatment effect heterogeneity

Alexandros Rekkas

Contents

1 Introduction **5**

Chapter 1

Introduction

Baseline risk is a crucial component of medical decision making. Because it provides personalized quantification of the likelihood for unwanted events, it is often used to guide treatment recommendations. For example in the European Society of Cardiology and the European Society of Hypertension guidelines of 2018 for the management of arterial hypertension, treatment initiation is based—among other things—on the patient's baseline 10-year cardiovascular risk [https://doi.org/10.1093/eurheartj/ehy339]. Similarly, an algorithm for the treatment of osteoporosis has been suggested, based on the stratification of patients on their 10-year hip or major osteoporotic fracture risk [https://doi.org/10.1007/s00198-019-05176-3].

Best practices for developing prediction models, evaluating their performance, and guiding their application in practice have been central focus of methodological research [refs]. Usually, prediction models are evaluated on their discriminative performance, i.e. their ability to separate lower from higher risk patients, and their calibration, i.e. the agreement of predicted risk to observed event rates [Ewout book]. These measures—though they are useful for assessing the predictive performance of one or multiple candidate prediction models—are not informative when it comes to applying these models in practice. Baseline risk is only one of the crucial pieces required for predicting individual responses to treatment. Knowledge of the patients' responsiveness to treatment, their vulnerability to side-effects and their utilities for other relevant outcomes is necessary information required for making truly informed clinical decisions [Kravitz, 2004]. Our aim was to explore approaches that incorporate baseline risk as the basis for medical decision making, shifting the focus from outcome prediction to the evaluation of treatment effect heterogeneity.

In order to provide the most optimal medical care, doctors are advised to align their clinical practice with the results of well-conducted clinical trials, or the aggregated results from multiple such trials [Greenfield 2007]. This approach implicitly assumes that all patients eligible for treatment experience the same

effects (benefits and harms) of treatment as the reference trial population. However, the estimated treatment effect is often an average of heterogeneous treatment effects—treatment effects vary across patients—and, as such, may not be applicable to most patient subgroups, let alone individual patients. If a treatment causes a serious adverse event, then treating all patients on the basis of an observed overall positive treatment effect may be harmful for some [Rothwell, Lancet 1995; …].

Conceptually, heterogeneity of treatment effect (HTE) is the variation of treatment effects on the individual level within the population [https://doi.org/10.1111/j.0887-378X.2004.00327.x]. The identification and quantification of HTE 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; [Holland, 1986]).

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. More individualized treatment effects are derived from the average effects estimated within a subgroup of similar patients. However, patient similarity is not straightforward to assess. Patients differ in a vast number of characterstics which 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 of false positives introduced from multiple testing.

Baseline risk, i.e. the probability—given 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 the treatment effect size. Low risk patients can only experience minimal treatment benefit before their risk is reduced to zero, while high risk patients can benefit much more. 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].

Baseline risk can also be used for directly predicting individual treatment benefit [Califf; Dahabreh, IJE 2016]. For example Califf et al [ref] predicted individual benefits regarding mortality with tissue plasminogen activator (tPA) compared to streptokinase treatment in patients with acute myocardial infarction using baseline mortality risk and assuming a constant relative tPA treatment effect. However, relative treatment effect does not need to be assumed constant. Modeling more flexible interactions of treatment with baseline outcome risk may provide more informative absolute benefit predictions for individual patients.

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 effect modifiers are known and the available sample size provides adequate statistical power for evaluating treatment-covariate interactions, modeling these interactions would be the optimal approach for assessing HTE. However, this approach may lead to overfitting and unstable estimates for the interaction effects [Balan, JCE].

Healthcare data is routinely collected by general practitioners, hospitals, insurance companies, and many other private or public bodies and 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 even a single such database. However, as this data is not being accumulated for research purposes, it suffers from many biases causing many commonly used methods to fail. Doctors prescribing a specific treatment expect—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 challenging.

If all relevant patient characteristics on which the treating physician based their decision have been captured in the observational dataset, methods are available 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-specific 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 between patients receiving the treatment under study and those who are not. Consequently, patients within this subset can be assumed to be randomized.

Unfortunately, not all the information that was used to decide on treatment is captured. As a consequence, propensity score adjustment will not suffice to evaluate treatment effects using the observational data, be it overall or sub-

group effects. Sensitivity analyses searching for evidence of this systematic unmeasured imbalances have been proposed and can be of assistance in many situations [refs].

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. In addition, more fundamental disparities between databases also factor in database incopatibility: different types of information are recorded in different databases. 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. 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 the 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 high level of standardization and scalability for observational studies was achieved. 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 the network providing effect estimates derived in different populations. The fragmented information scattered across multiple databases can now be summarized in a consistent way to give a fuller picture.

The power of the common database structure was demonstrated in a large-scale comparative effectiveness study of first-line treatment for hypertension [Suchard, Lancet 2019]. This study compared five different first-line drug classes prescribed for hypertension regarding three primary effectiveness, six secondary effectiveness, and 46 safety outcomes across a global network of 9 observational databases, all mapped to OMOP-CDM. A framework following best practices for carrying out such analyses was proposed and implemented on a large scale. The results complemented the already available evidence generated in clinical trials, confirming earlier findings and providing effect estimates on previously unexplored comparisons.

Observational databases provide access to massive numbers of “real-life” patients. This motivates the exploration of methods for the assessment of treatment effect heterogeneity in the observational setting despite the challenges inherent to this type of data. The statistical power problems related to subgroup analyses can still be present, as observational data is high-dimensional, i.e. the number of measured patient characteristics increases with the number of patients. Attempting a treatment effect modeling approach, where treatment-covariate interactions are modeled for the prediction of individualized treatment

benefits, suffers from the same statistical power issues and often results to highly variable estimates. Therefore, using baseline outcome risk as the subgrouping variable, can provide good insight of treatment effect heterogeneity in the observational setting, as well. Modern libraries for developing risk prediction models are available and—capitalizing on OMOP-CDM—can be easily applied across databases with millions of patients.

The overall aim of this thesis is to explore approaches that incorporate baseline risk as the basis for medical decision making, shifting the focus from outcome prediction to the evaluation treatment effect heterogeneity. We will explore methods and applications in both the clinical trial and the observational setting. More specifically, the main research questions are:

- Aim 1: *How to develop and present a prediction model to be used in clinical practice?* We will explore the development of multi-purpose prediction models, i.e. models that can be used to predict patient risks for multiple outcomes. We will also demonstrate different approaches for their application in practice.
- Aim 2: *How to use risk estimates from a prediction model to assess treatment effect heterogeneity?* We will explore the literature for methods on the assessment of treatment effect heterogeneity. We will develop and apply a framework for risk-based assessment of treatment effect heterogeneity in the observational setting. Finally, we will explore methods for making individualized risk-based benefit predictions.

In **Chapter 2** we develop a model for the prediction of 5-year recurrence risk in sentinel node positive melanoma patients, using data from nine European Organization for Research and Treatment of Cancer centers. We calibrate the recurrence model to predict 5-year risk of distant metastasis and overall mortality and develop a nomogram for graphical representation. The models are, then, extenrally validated.

In **Chapter 3** we develop a model for the prediction of 28-day mortality for patients presenting at the emergency department with suspected COVID-19 infection at four large Dutch hospitals between March and August, 2020. We predict 28-day admission to the intensive care unit by calibrating the mortality model. An easy to use web application is also supplied. We perform temporal validation to assess model performance using data between September and December, 2020.

In **Chapter 4** we present the results of a scoping literature review of regression modeling approaches for the assessment of treatment effect heterogeneity in the clinical trial setting. The identified methods are divided into broader categories based on how they incorporate prognostic factors and treatment effect modifiers.

In **Chapter 5** we develop a standardized scalable framework for the assessment of treatment effect heterogeneity using a risk-stratified approach in the observational setting. We, also, develop the software for the execution of the framework

in observational databases mapped to OMOP-CDM. We, finally, demonstrate the application of the framework, assessing treatment effect heterogeneity in first-line treatment for hypertension across three US claims databases.

In **Chapter 6** we apply the standardized framework to evaluate effect heterogeneity of teripatide treatment compared to oral bisphosphonates in female patients above the age of 50 with established osteoporosis. We use different risk stratification approaches based on quantiles of predicted risk and externally derived risk thresholds for treatment. We evaluate the presence of residual confounding using sensitivity analyses.

In **Chapter 7** we compare different risk-based methods for predicting individualized treatment effects using an extensive simulation study. We only consider the clinical trial setting were treatment is admistered at random.

Finally, in **Chapter 8** we present a general discussion along with perspectives on future work.