Journal of Data Science 19 (4), 519–527
October 2021
DOI: 10.6339/21-JDS1029
Philosophies of Data Science

So You Developed a Clinical Prediction Model, Now What?

Jaime Lynn Speiser[1,*]

[1]*Department of Biostatistics and Data Science, Wake Forest School of Medicine, Winston-Salem NC, USA*

Abstract

A recent trend in medical research is to develop prediction models aiming to improve patient care and health outcomes. While statisticians and data scientists are well-trained in the methods and process of developing a prediction model, their role post-model-development is less clear. This paper covers the critical scientific reasoning step in the prediction pipeline after a model is developed. Working collaboratively with domain experts, statisticians and data scientists should critically evaluate models, carefully implement models into practice, and assess the model's impact in real world settings. Constructs from implementation science are discussed in the context of prediction modeling. The paper focuses on clinical prediction models, but these ideas apply to other domains as well.

1 Introduction

A recent trend in medical research is to develop prediction models aiming to improve patient care and health outcomes. Model development typically involves an interdisciplinary team of researchers, including domain experts in medicine and/or clinical practice, as well as statisticians or data scientists. Once a team is established, the typical model development process may be as follows.

Researchers work together to identify an outcome of interest, predictors available to ascertain the outcome, and a suitable dataset containing these variables used to conduct analysis (Royston et al., 2009). There is a plethora of models available, ranging from traditional regression to newer machine learning approaches such as neural networks. Often, many models are compiled and compared. An optimal model is selected based on pre-defined performance criteria and validation of some form (e.g., split sampling or cross-validation). If a comparable dataset is available, the model is evaluated with external data to assess performance and the generalizability of results (Altman et al., 2009). Assuming all goes well with the development process, the prediction model is then released into the wild, usually through a publication in a journal. This completes a typical model development process.

This typical process describes the initial development of a clinical prediction model, but developing a model is only the beginning of the pipeline. After a prediction model is developed, there are many other aspects that need to be addressed, including critical evaluation of the model, implementation of the model into practice, and evaluation of how the model is being used by the intended audience. As statisticians and data scientists, we are well-trained in the methods and process of developing a prediction model, but we have less experience with what comes next. This paper will address key steps that should be taken after a prediction model is developed, including critical evaluation of the data and model, and implementation of the model

* Email: jspeiser@wakehealth.edu.

Received October 14, 2021; Accepted October 25, 2021

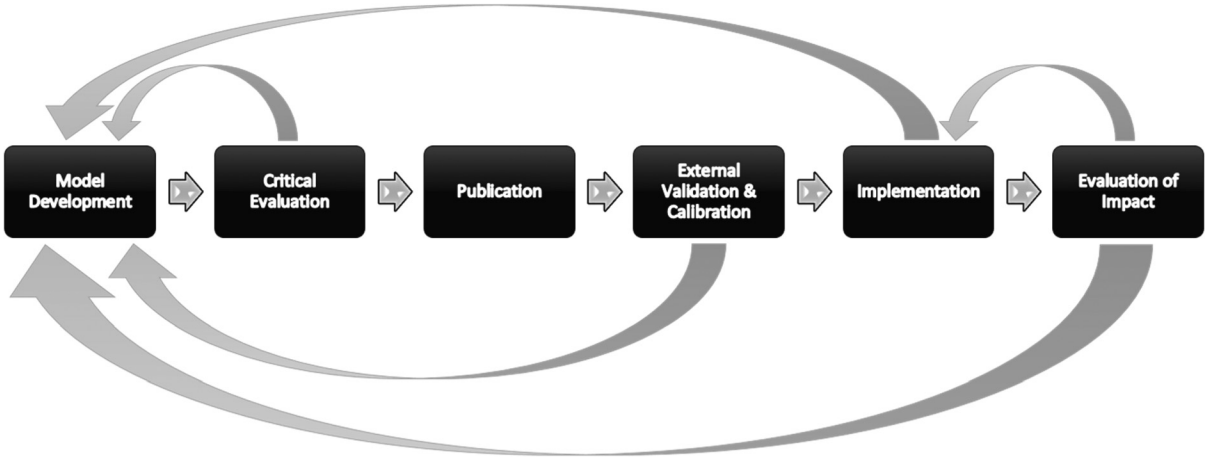

Figure 1: Prediction Model Process after Model Development.

into clinical practice. The paper will focus on clinical prediction models, but these ideas apply to other domains as well.

The framework described in this paper is depicted in Figure 1. The main process after model development includes critical evaluation, publication, external validation and calibration, implementation, and evaluation of impact. In an ideal world, these would occur in a linear fashion in this order; however, in practice it is often necessary to iterate over previous steps depending on discoveries from the current step. These are represented by grey arrows linking the steps within Figure 1. This paper will discuss each of these steps. In this paper, we focus on prediction modeling in an academic research setting. This process may look different in industry, but critical evaluation and careful implementation into practice remain essential for the prediction model pipeline.

2 Critical Evaluation of the Prediction Model

Critical evaluation of a prediction model can and should take many forms. Of course, the typical assessment of performance characteristics should be conducted (e.g., accuracy assessment via area under the receiver operating curve, mean squared error, etc.), but there are other ways that models must be evaluated from a critical scientific perspective. As an English professor might deconstruct a sentence, statisticians and data scientists should deconstruct elements of prediction models to critically evaluate its components, both individually and holistically. This evaluation is a collaborative effort between statisticians and data scientists along with clinical domain experts (e.g., clinicians and medical providers).

2.1 Evaluating the Outcome Variable

We can begin by considering the outcome variable. What is the main outcome of clinical interest? Does this align with the outcome variable that is available within the dataset? Is the outcome variable identical to the clinical goal, or is it a proxy for something else that is more important?

There are many examples in medicine where researchers are interested in one outcome, but develop the prediction model on a proxy variable. One example that I have encountered is the

prediction of quality of life in older adults (Speiser et al., 2021). The outcome that is really of interest is independence of older adults, but this is a somewhat nebulous concept. In practice, we use outcomes that are more easily defined and measured, such as the ability to walk for 100 meters or walk up a flight of stairs. Proxy outcome measures are commonplace for quality control studies in a healthcare system. For example, false-positive alarms in patient monitoring is often used as a proxy for alarm fatigue of clinical staff. Also, early-warning time of a clinical alarm is a proxy for additional time for clinical intervention. The gap between a proxy variable and the true outcome of interest may not be readily apparent. It may require further analysis to assess. No matter the challenge, this gap should be kept in mind when developing prediction models.

Sometimes outcomes are measured or collected in ways that make modeling and interpretation challenging. For example, I worked on a number of studies about acute liver failure, a rare and devastating condition characterized by rapid deterioration of the liver. A major question that arises is which patients should get a liver transplant (Speiser et al., 2015). The main outcome of interest is mortality, but in the presence of transplant, prediction modeling is challenging because mortality and transplantation are dependent events. A patient's mortality status cannot be observed without having the transplant. This presents a unique problem for prediction modeling because it is not straightforward which outcome should be used. Many researchers use a composite outcome consisting of all "bad" outcomes (death and transplant) and make the assumption that patients requiring a transplant would have died without one.

When developing clinical prediction models, it is essential to think through how the prediction models will be used and interpreted when deciding on which outcome variable the model should be trained or optimized. The statistician or data scientist has many options when discovering misalignment between the outcome variable and the clinical goal. Revising the outcome variable is one option. Another is to revise the cost function against which the model is evaluated.

2.2 Evaluating the Predictors

Aside from the outcome being modeled, the model's predictors should be critically assessed. Are the correct predictors included in models? This seems like a simple question, and it is, but it is an essential question to pose. Deciding on the predictors in the model can be driven by automated processes, such as variable selection, but this also should be evaluated by content area experts to ensure that variables included in the model make sense from a biological and practical perspective. Following assessment of variables included in the models, we should also ask: are there any features that should be included in the model that are not (e.g., unmeasured but important variables or latent variables)? If so, are there proxy variables available for these? Thoroughly evaluating each variable in the model, as well as those we would like to include in the model, is a first step in critically analyzing the predictors.

Having considered predictors individually, the associations between predictors and outcomes should be considered next. Do the associations between predictors and outcome align with clinical intuition or knowledge? The direction and strength of associations between predictors and outcome ideally should be consistent with clinical observation. This is important for all types of models, but is especially key for machine learning models that are prone to overfitting. Assessing the relationship between predictors and outcome will differ based on the method used to develop the model. For instance, when using decision trees, I print out the tree diagram and go through each node of the tree with clinical collaborators to make sure that the thresholds are consistent with their observations. For more complex machine learning methods, variable

importance measures and partial dependence plots are often helpful for delving into these associations. Especially in a medical setting, it is not enough to have a prediction model perform well in terms of accuracy; the models also need to be biologically plausible.

It can be challenging to determine whether the associations between predictors and outcome are plausible. Models usually contain a multitude of predictors that may interact together, but clinical literature may have only assessed each predictor with respect to a single dimension, or in control cases with little confounding. Alternatively, clinical literature may be incomplete with respect to currently-available data, and machine learning techniques may have been used specifically to handle data channels that clinicians cannot (e.g., high throughput data). An example of this is prediction modeling for patient triage in critical care units, where patients are continuously monitored by digital devices during their hospital stay. A statistician or data scientist may plot and assess hundreds of hours of a patient's data, whereas a clinician may have decades of in-person observations for previous patients. Ideally, associations between predictors and outcome for models will agree with clinical observations, but there may be cases where these disagree. When this happens, clinical experts and statisticians or data scientists should work together to determine what might have caused these differences, and if anything should be done to address this in the model. For example, if nodes of decision trees do not make sense, pruning could be applied to make the model more consistent with clinical observation. Although challenging in some scenarios, it is important for the predictors included in models to be critically assessed in the clinical context.

2.3 Evaluating the Model as a Whole

After critically analyzing the inputs and outputs of the prediction model, one should consider model calibration and the cost function against which it is trained. Is model performance optimal for the clinical application? For example, a model could perform well according to the c-statistic (or area under the receiver operating curve), but its sensitivity may be low. Differing costs of incorrect predictions need to be considered when developing prediction models. A screening tool used to identify patients who may benefit from additional testing may have high sensitivity and low specificity. This is acceptable because further testing would correctly identify the outcome. However, if the clinical scenario has an equal cost of incorrect predictions, then the model or its cost function should aim to balance sensitivity and specificity measures. Figuring out how to calibrate the model based on the cost of incorrect predictions is a process that should involve collaboration with clinical experts. In particular, for external validation, there are many calibration techniques that can be applied to boost performance in new datasets (Moons et al., 2012). For example, one might keep the same basic prediction model form, but add a random intercept that adjusts for different sites.

In addition to calibration of the cost function, a key evaluation of the model as a whole involves comparing it to existing prediction models. Are there other prediction models for the same or similar outcomes already available and in use? If so, how does this prediction model augment what is already available? This augmentation could be in the form of a better performing model or a model that is simpler to use in practice (e.g., with less predictor variables or variables that are readily available, such as electronic health record (EHR) data). Alternatively, the new model could detect or capitalize on physiology or mechanisms that are neglected by previous models. For a prediction model to be adopted into practice, it is essential that it provides some benefit over existing models.

3 Implementation of the Prediction Model

Ideally, the critical evaluation process described in Section 2 will take place before publication (Figure 1). Section 3 focuses on implementation of the prediction model. In academic research settings, this typically happens after a prediction model is published in a journal. Although many papers present prediction models for various outcomes in medicine, there are fewer papers that focus specifically on how the model is incorporated into practice and how the model is used by the intended audience. For example, a systematic review by de Pablo and colleagues found that only 0.2% (1 out of 584 studies) of prediction models for psychiatry outcomes had been implemented into practice (Salazar de Pablo et al., 2021). While not all prediction models should be incorporated into practice (i.e., models with poor performance or models that have not been robustly validated (Kansagara et al., 2011; Belsher et al., 2019)), there are many high-performing prediction models that are never used by the intended audience. To address this translational gap, we can draw from the field of implementation science.

3.1 Implementation Science and Clinical Prediction Modeling

Implementation science is a relatively new area in medical research over the past decade that focuses on how results from research studies, such as clinical trials, are incorporated into practice in real world settings (Glasgow et al., 2012). This typically involves disseminating the research finding as well as evaluating its effectiveness or impact. The motivation behind implementation science is that interventions and findings from research studies do not always translate into improved care or outcomes for patients in real world settings. Implementation science provides theory, methods, and practical guidance about translating medical findings into practice (Rapport et al., 2018; Bauer and Kirchner, 2020). By studying how interventions or findings are best implemented into practice and then evaluating how this actually impacts health outcomes, this field is an essential part of the translational science pipeline.

There is great opportunity for statisticians and data scientists involved in prediction modeling to use constructs from implementation science for disseminating and evaluating models in a practical setting. For example, Kappen and colleagues discuss implementation science for clinical prediction models and give practical advice about how to design studies that evaluate prediction models in practice (Kappen et al., 2018). In the next sections of the paper, dissemination of prediction models will be discussed through an implementation science lens, as well as challenges that may arise with prediction model implementation.

3.2 Model Dissemination

The first step in implementing a prediction model is figuring out how to produce the model and deliver it to the end users (i.e., often medical professionals, for clinical prediction modeling). Dissemination of a model can take many forms: some models can be written out as an equation (e.g. regression), others are more complex and require an online application with software running in the background (e.g. random forests or neural networks). Some simple regression models with a large number of engineered features may require an online application as well (e.g., the Acute Physiology and Chronic Health Evaluation (APACHE) II score for predicting hospital mortality (Knaus et al., 1985)). It is essential to think about the end users of the model and their workflow. Often statisticians and data scientists disseminate their work on repositories such as GitHub or share their code along with a publication, which is great for quantitatively-trained researchers; however, medical professionals are not to be able to access or use such code in practice within the

confines of clinical visits. In medicine, an optimal dissemination method may be to incorporate prediction models into the EHR system because it is less disruptive to clinical work flow and requires less effort for the end user. A less-effortful option than incorporation in the EHR is to develop an online application, such as an R Shiny program. This would allow users to access the prediction model through a webpage with a fillable form to input variables.

It is critical to work with clinical collaborators and people who will actually be using the prediction model in practice to figure out how to make it assessable, understandable, and efficient. A lot of thought and discussion should focus on deciding how to present prediction model results. For example, are risk scores such as probabilities useful, or would the users prefer a warning flag that converts a risk score into a binary or categorical outcome? Should measures of uncertainty be presented, and if so, how? Kompa and colleagues provide an excellent article about uncertainty in medical machine learning and conclude that models should include the ability to conclude "I'm not sure" or "I don't know" when predictions are ambiguous (Kompa et al., 2021). Ideally, prediction model results should give the users the information they need and should not contain superfluous results that may be distracting or confusing. If the end users do not understand the model output, then they will likely not use it in a meaningful way to improve patient care or outcomes.

3.3 Challenges with Implementation

There are many challenges to implementing clinical prediction models into practice. A major challenge is missing data. Predictor variables used in models and trained on retrospective datasets are often not available in real world settings. A common example of this is EHR data, which often has unknown or incorrect information. Missing EHR data is often missing not at random (MNAR), meaning that its missingness is associated with the outcome. For instance, a model predicting hospital admission for drug overdose may use illegal drug use as a predictor. This data may be MNAR because people who use illegal drugs might not disclose this information. Thus, missing values for illegal drug use may be associated with higher likelihood of hospital admission. Additionally, when data is collected repeatedly or in a time series, missingness is a challenge. Many prediction models require complete data in order to make predictions, so this presents a barrier to implementation in practice. Although imputation may be used to fill in missing values, this introduces more error in prediction (i.e., imputation accuracy) and may violate assumptions (i.e., missing completely at random or missing at random patterns). The problem of missing data also relates back to the critical evaluation of predictors discussed in the previous section, because statisticians and data scientists should consider the likelihood of variables being non-missing in real world data in order to include them in modeling. Critical judgement may be required to decide the tradeoff between including features and the challenges of missingness.

Administrative or legal issues are other challenges in implementing prediction models into practice. Returning to the example of trying to implement a prediction model into the EHR, researchers may run into administrative hurdles because it is hard to get new features programmed in the EHR and adopted by the end users. Getting new features into the EHR may be challenging because of the balance between not overwhelming medical professionals with information versus providing helpful features that can improve care. When multiple researchers are vying to get their features included in the EHR, this creates an administrative conundrum. Furthermore, incorporating models into the EHR is very costly due to licensing, programming, and server fees. On top of that, legal issues may arise, depending on the outcome and predictors in the model.

A legal team may need to write a disclaimer for a prediction model so that the researchers who developed it are protected. There may also be issues with data privacy and anonymity of data, particularly in medicine because of the Health Insurance Portability and Accountability Act and other protections for healthcare data. Missing data, administrative challenges, financial burden, and legal considerations are a few barriers that may be encountered when trying to implement prediction models into practice. The more that a researcher understands and can work within these constraints, the more likely their model is to see real-world use.

3.4 Evaluate How Clinicians Are Using the Predictions in Practice

After a prediction model is implemented into practice, it should be evaluated to ensure that it is practical and impactful. The first evaluations that should be conducted involve assessing the feasibility of the model in practice. How burdensome is data collection of input variables to clinical staff? Is it feasible to collect all input variables within the confines of clinical visits? A common viewpoint of statisticians and data scientists is "more data is better," but in practical settings, this can stymie model implementation and use. If the burden of data collection is too great, then the project is unlikely to be sustainable for the future. Statisticians and data scientists should know whether their model can be sustained, because this will help them develop models that are useful in real world settings. If the work is not sustainable in the long run, statisticians and data scientists may develop innovative technical solutions that would facilitate use of the prediction model. For example, if some input variables are costly to obtain, alternative methods could be investigated that would circumvent using those variables. Although statisticians and data scientists have the potential to develop some really interesting, complex models, if they are not practical then the endeavor is futile.

If a prediction model is practical to use in real world settings, the next step is to assess the impact of its use (Moons et al., 2009). There are many types of prediction models, and the meaning of "impact" will vary in different scenarios. A good starting place for evaluating impact is asking questions to the end users of the model. In many clinical prediction model studies, the end users are clinicians and medical professionals. How do clinicians perceive the prediction for a patient's outcome? Do they agree or disagree with the algorithm's prediction? Was it surprising or something obvious? In my experience, clinicians say that extreme cases are easy to predict, but cases "in the middle" are a lot more difficult. Clinicians are often looking for guidance on these "middle" cases. Assessing how clinicians and end users of models perceive predictions in a formal study helps determine how the models might be used, or why they would not be used, in practice.

Perhaps the two most important questions that need to be asked are: 1) did the prediction model impact clinical decisions about care and 2) did those clinical decisions impact health outcomes for the patient? At the end of the day, if a well-performing prediction model is not augmenting the current process or outcomes for the patient, it is useless.

3.5 Shadow Domain Experts

Before concluding the paper, I will take a brief aside to discuss the value in thoroughly understanding the data used to develop prediction models and how it is collected. I highly recommend shadowing a clinician during patient visits. I have done this in two different clinical settings, with a hepatologist in a liver transplant clinic and a primary care provider in a geriatrics clinic. It was an invaluable experience to see how the data is collected and how prediction models that

I develop may fit into the existing setup to augment clinician decisions and patient care. As a quantitatively-trained researcher, I never thought I would actually see patients and talk to them, but I learned so much about where the data comes from and who the patients are. After shadowing clinicians, patients became real to me, not just lines in a dataset. Not only have I learned a lot about the clinical areas that benefit my research, I also gained a level of personal meaning in the work that motivates me to help improve care and health outcomes. Regardless of the application area, shadowing a domain expert is beneficial for understanding the data, and I highly recommend taking the time to dig into the data as it is collected.

4 Conclusion

For projects involving prediction models, critical evaluation and careful implementation are necessary to ensure that models will be appropriate and useful in practice. Although statisticians and data scientists developing medical prediction models are often focused on technical details, they should also keep the big picture in mind: the overall goal is to improve clinical care and patient outcomes. A prediction model with good performance is great, but it does not have impact if it is not available to clinicians in a way that is useful and will augment their knowledge to make clinical decisions. Developing a prediction model is just the start of the process, and it must be followed by rigorous scientific evaluation, implementation into practice, and evaluation of impact.

Funding

This work was supported by a K25 Career Development Grant from the National Institute on Aging (K25AG068253). The views expressed are those of the author, not of the funding agency.

References

Royston P, Moons KG, Altman DG, Vergouwe Y (2009). Prognosis and prognostic research: developing a prognostic model. *BMJ*, 338:b604.

Altman DG, Vergouwe Y, Royston P, Moons KG (2009). Prognosis and prognostic research: validating a prognostic model. *BMJ*, 338:b605.

Speiser JL, Callahan KE, Ip EH, et al. (2021). Predicting future mobility limitation in older adults: A machine learning analysis of health ABC study data. *The Journals of Gerontology: Series A*.

Speiser JL, Lee WM, Karvellas CJ, Group USALFS (2015). Predicting outcome on admission and post-admission for acetaminophen-induced acute liver failure using classification and regression tree models. *PLoS One*, 10(4): e0122929.

Moons KG, Kengne AP, Grobbee DE, et al. (2012). Risk prediction models: II. External validation, model updating, and impact assessment. *Heart*, 98(9): 691–698.

Salazar de Pablo G, Studerus E, Vaquerizo-Serrano J, et al. (2021). Implementing precision psychiatry: A systematic review of individualized prediction models for clinical practice. *Schizophrenia Bulletin*, 47(2): 284–297.

Kansagara D, Englander H, Salanitro A, et al. (2011). Risk prediction models for hospital readmission: A systematic review. *JAMA*, 306(15): 1688–1698.

Belsher BE, Smolenski DJ, Pruitt LD, et al. (2019). Prediction models for suicide attempts and deaths: A systematic review and simulation. *JAMA Psychiatry*, 76(6): 642–651.

Glasgow RE, Vinson C, Chambers D, Khoury MJ, Kaplan RM, National HC (2012). Institutes of Health approaches to dissemination and implementation science: Current and future directions. *American Journal of Public Health*, 102(7): 1274–1281.

Rapport F, Clay-Williams R, Churruca K, Shih P, Hogden A, Braithwaite J (2018). The struggle of translating science into action: Foundational concepts of implementation science. *Journal of Evaluation in Clinical Practice*, 24(1): 117–126.

Bauer MS, Kirchner J (2020). Implementation science: What is it and why should I care? *Psychiatry Research*, 283: 112376.

Kappen TH, van Klei WA, van Wolfswinkel L, Kalkman CJ, Vergouwe Y, Moons KG (2018). Evaluating the impact of prediction models: Lessons learned, challenges, and recommendations. *Diagnostic and Prognostic Research*, 2(1): 1–11.

Knaus WA, Draper EA, Wagner DP, Zimmerman JE (1985). APACHE II: A severity of disease classification system. *Critical Care Medicine*, 13(10): 818–829.

Kompa B, Snoek J, Beam AL (2021). Second opinion needed: Communicating uncertainty in medical machine learning. *NPJ Digital Medicine*, 4(1): 1–6.

Moons KG, Altman DG, Vergouwe Y, Royston P (2009). Prognosis and prognostic research: Application and impact of prognostic models in clinical practice. *BMJ*, 338: b606.