Clinical utility

Michael C Sachs

SMARTbiomed
Pioneer Centre

Why prediction?

Context: We wish to know an unknown or future event

- Underlying disease state (diagnosis/classification)
- Future disease outcome (prognosis)
- Response to treatment

Step 1: Form predictions based on observations

- Medical tests, Questionnaires
- Genetic mutations/expression, register data

Step 2: Assess the value of the prediction model

- Accuracy?
- Utility for determining treatment?

The value of a prediction model

What is the right way to assess the value of a prediction model?
Answer: It depends. What is the intended use?

- Implement a new policy or screening program on a population level
- Guide treatments for individual patients
- Allocate funds for further research and development?

Use cases for clinical prediction models

Intended Use	**Examples**
Diagnose	eGFR, cardiac monitors
Determine treatment	HER2, Mammaprint, OncotypeDX
Inform decisions	Framingham, SCORE2
Research only	CCI, Inflammatory burden score

- High risk use requires a high level of rigor and high quality evidence that using the model benefits patients, on average, compared to the standard of care (clinical utility, effectiveness).

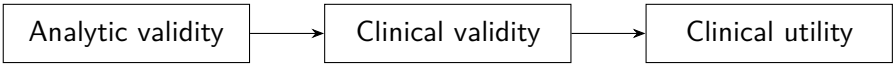
The leaky pipe

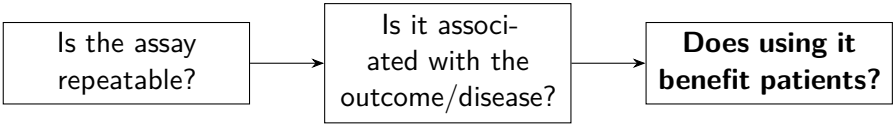
Figure 1

Visualization of the difficulties from development to successful deployment of prediction models.
Reproduced from van Royen et al. (2022), with permission of the European Respiratory Society (ERS) (2025).

Terminology from cancer biomarker research

The progression from the lab to clinical practice





Targets for evaluation

- Prediction models generally target validity, i.e., find a transformation of covariates that is strongly associated with the outcome
- Many of the common performance metrics only address validity, e.g., MSE, calibration, C-statistic
- Causal prediction models target interventional versions of these

In our paper "Aim for clinical utility, not just predictive accuracy" we argue that clinical utility can and should be targeted directly in the evaluation and maybe even in the development of prediction models [Sachs et al., 2020].

Key features of clinical utility

- **Outcomes** - health related outcomes that matter to patients. Balances both benefits and risks.
- **Actions** - a clear space of possible actions taken in response to the model
- **Comparative** - relative to another clearly defined strategy or the standard of care
- **Implementable** - based on things that are measurable in the clinic, able to be used in the appropriate timeframe, transparent, clear plan for updating, ...

Clearly we are talking about a *causal effect*: the effect of using the model compared to doing something else.

The statistical task of developing a model that is a good predictor of the outcome (high AUC) is not obviously linked to this goal.

The Estimand

$$E\{Y^{(\text{used prediction model})}\} - E\{Y^{(\text{used standard of care})}\}$$

- Y is an outcome that matters to patient
- "used" implies there is a strategy, e.g., treatment guidance
- Compared to the standard of care strategy
- Potential outcomes, we are interested in the causal effect of following the prediction model strategy

Example cont.

- **The strategies:** If the QRISK score is high, start taking statins, otherwise wait and see
- **The outcome:** Cardiovascular disease event within 10 years

How would you design the ideal study to estimate the clinical utility in this context?

Study designs

1. Direct randomized controlled trial with two arms: use the QRISK calculator to decide when to start statins versus the standard of care.
2. Advantages: easy to understand, estimand is clearly identified
3. Disadvantages: costly, time consuming, difficult to recruit participants

Study designs (2)

Rewrite the first term of the estimand as

$$Pr\{\text{low risk}\} * E\{Y^{(\text{no statins})} | \text{low risk}\} + Pr\{\text{high risk}\} * E\{Y^{(\text{statins})} | \text{high risk}\}$$

1. It can be estimated in a randomized trial of statins, if we can calculate the risk score. But the comparator depends on what the standard of care is.
2. It can be estimated in observational data if we measure all confounders of the statins $\rightarrow Y$ relationship.

Using an emulated trial in observational data

Same as Hernán and Robins [Hernán and Robins, 2016]

- Use the same inclusion/exclusion criteria that a clinical trial might
- Define a grace period for use of the deterministic decision rule
- Deal with people that die/have the event prior to the end of the grace period

Different than Hernán and Robins

- The estimand is $E\{Y^{\text{used prediction model}}\} - E\{Y^{\text{standard of care}}\}$
- since $E\{Y^{\text{standard of care}}\}$ is literally what we observe, this has no confounding
- $E\{Y^{\text{used prediction model}}\}$ can be decomposed to allow for estimation in the observed data

In general

Given a prediction model $\hat{m}(X)$ and a treatment decision rule of this form

$$\begin{aligned}&\text{use } a_1 \text{ if } \hat{m}(X) \in B_1\\ &\text{use } a_2 \text{ if } \hat{m}(X) \in B_2\\ &\dots\\ &\text{use } a_k \text{ if } \hat{m}(X) \in B_k\end{aligned}$$

for a partition $[B_1, \dots, B_k]$ of the output space, we can write the estimand $E\{Y^{\text{used prediction model}}\}$ as

$$\sum_{j=1}^{k} Pr\{\hat{m}(X) \in B_j\} E\{Y^{a_j} | \hat{m}(X) \in B_j\}.$$

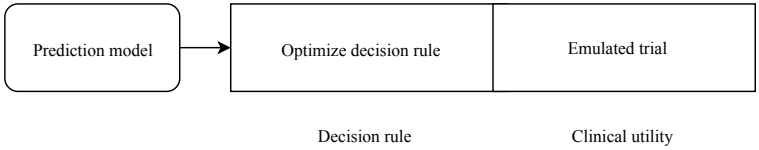
What data and assumptions do you need to estimate these?

Answer

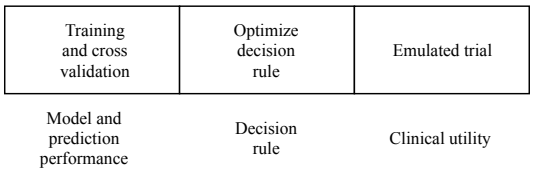
1. Data where X, Y, A are observed
2. Positivity holds, i.e., $P(A = a_j|X) > 0$ for all a_j, X
3. Conditional exchangeability given L and consistency

Then we can use g-computation, IPW, doubly robust methods, etc.

Optimizing the prediction model for utility



Uses existing prediction model and optimizes decision rule for that model. Evaluates clinical utility in independent data.



Identifies prediction based optimal decision rule, and evaluates its utility. Requires lots of data.

Optimizing for clinical utility

Let X be the vector of covariates under consideration for the model, A be the treatments/actions, and Y be the outcome.
Assuming people want to minimize their Y, the optimal action is for covariate vector x is

$$\text{argmin}_a\{E[Y^a|X = x]\},$$

i.e., the a that minimizes their risk.
So, we want a mapping from X to A, given data, output a treatment decision.
What data/assumptions do you need to do this?

Evaluating clinical decision support in observational data

- Maybe it is too ambitious to aim for a model that directs medical action from data
  - It may be considered a high-risk medical device subject to more scrutiny
  - We do not want to take the human out of the loop
- Instead aim for a clinical decision support, provide information and let the patient/doctor decide course of action
- How to evaluate the clinical utility of such a model?

Requires some data collection.

Conduct a survey

We need to know $Pr\{A = a_j|\hat{m}(X)\}$, for each possible action a_j.

- Compile a list of scenarios (covariate vectors) and their model output $\hat{m}(X)$.
- Ask some (a random sample) doctors what action a_j from among a list of actions they would take when presented with the information.
- Estimate probability of each action for a series of covariate vectors.

Treat the use of the decision support as a stochastic intervention [Haneuse and Rotnitzky, 2013].

Example

Two scenarios and two treatments.

$\hat{m}(X)$	$\hat{P}(A=1)$	$\hat{P}(A=0)$
$< .5$	.3	.7
$\geq .5$	.9	.1

Probabilities are estimated as the proportion of clinicians who recommend that action when presented with that information (or subjective probabilities).
The clinical utility of the decision support system is that of the conditional treatment policy described in the table.
This can feasibly be estimated if $E[Y^a|\hat{m}(X)]$ can be and then compared to the standard of care.

A future coding exercise

Using the steno synthetic data,

- develop a model that takes in covariates and outputs *a treatment decision*
- Apply the model to the deployment data, and send me the results

To score the performance, I would

- Generate observations by *intervening* according to the suggested treatment by your model
- Compare the expected outcomes to the outcomes under the standard of care regime

Again, this is difficult to do in real life, but it illustrates what you are trying to get at with clinical utility

Summary

- Clinical utility of a prediction model is like efficacy of a drug
- Whether and how you can estimate this depends on the intended use of the model
- Ideal: an RCT
- Often feasible: observational data, maybe with some new data collection
- Always feasible: synthetic data, does your method work under your assumptions?
- Provides numeric evidence to support statements like "further research is needed to get this into practice" or "this model is not promising and should be abandoned in favor of something else" where accuracy measures do not

Conclusion

- It is not a reasonable PhD project to get a new drug developed and into market; nor should it be expected to do the same for a prediction model
- We do not expect researchers to run hundreds of randomized prediction trials
- ... but we need a way to prioritize models for implementation
- A prediction model does not need to be causal to be useful (but it probably helps)
  - does using the prediction model improve outcomes?
  - *what if* questions about starting a treatment or doing some other intervention
- Lots of open research questions in this area

References I

Sebastian Haneuse and Andrea Rotnitzky. Estimation of the effect of interventions that modify the received treatment. *Statistics in medicine*, 32(30):5260–5277, 2013.

Miguel A Hernán and James M Robins. Using big data to emulate a target trial when a randomized trial is not available. *American journal of epidemiology*, 183(8):758–764, 2016.

Michael C. Sachs, Arvid Sjölander, and Erin E. Gabriel. Aim for Clinical Utility, Not Just Predictive Accuracy. *Epidemiology*, 31(3), 2020. ISSN 1044-3983. URL `https://journals.lww.com/epidem/Fulltext/2020/05000/Aim_for_Clinical_Utility,_Not_Just_Predictive.8.aspx`.