Author’s response to reviews

Title: Evaluating the impact of prediction models: lessons learned, challenges and recommendations

Authors:

Teus Kappen (t.kappen@umcutrecht.nl)

Wilton van Klei (W.A.vanKlei@umcutrecht.nl)

Leo van Wolswinkel (l.vanwolfswinkel@umcutrecht.nl)

Cor Kalkman (c.j.kalkman@umcutrecht.nl)

Yvonne Vergouwe (y.vergouwe@erasmusmc.nl)

Karel Moons (k.g.m.moons@umcutrecht.nl)

Version: 1 Date: 17 Mar 2018

Author’s response to reviews:

Reviewer #1:

1. The English is rather tortuous - suggest editing to improve readability

We thank the reviewer for pointing that out to us. It is sometimes easy to lose track of the complexity of one’s own work. We changed several of those tortuous sentences and paragraphs throughout the text.

2. The model doesn’t seem to have very good discrimination - could you write a sentence to either explain why this was not felt to be a problem (or explain to me if I have misunderstood).

We agree with the reviewer that the model doesn’t have very good discrimination. At the time, we were very pleased our model was comparable to other prediction models – and even a little better in a Dutch population. In addition, the calibration was very good. As prophylactic treatment would consist of choosing a number of interventions, a properly calibrated prediction model – even with a moderate C-statistic – would allow us and the physicians to identify specific risk groups and increase the number of interventions according to the predicted risks. We included our views at the time in the text.

Page 7, lines 4-8. “At the time, we considered the discriminative ability to be adequate with a C-statistic of 0.68, because other models had proven to have similar performance. Nonetheless, even with a moderate discriminative ability, a properly calibrated model would enable physicians


to identify specific risk groups of patients and increase the number of prophylactic interventions according to the predicted risks.”

3. It is always a problem to know if the association between predictors and outcomes changes over time and populations and practice changes; does this need more cycles of testing and validation? How often should this be done?

We thank the reviewer for making this excellent point. However, it is currently little information on how many cycles of testing and validation should be performed before a prediction model is ready for use in clinical practice. We only know that at least one external validation would be recommended. Prompted by the comments of the reviewer we adjusted the text accordingly.

Page 6, lines 22-27. “This has not yet resulted in a clear set of validation guidelines on when model performance will be adequate in the new setting or when further tailoring of the model is necessary. Until clear guidelines exist regarding the number of external validations that are needed before use in daily practice, we recommend to first quickly evaluate the need for customization of the model to the new setting in which the model will be implemented, even when a model has been thoroughly validated in other settings.”

4. If a model is ready for practice, it's surprising that predictors should be missing; is this because users didn't want to fill in the model, or because they thought it was irrelevant?

We agree that it would make sense to have all predictors available, as one would expect that a healthcare provider would want to have the best possible estimate. However, sometimes predictor variables are missed or incorrectly documented, especially when providers are simultaneously performing other tasks. Predictor variables may in the moment be difficult or even impossible to measure, e.g. when a device is not available. Automatically calculated models typically rely on routine data and are thus more prone to missingness. Whatever the reason for missing, it is important to impute the missing value rather than leaving it out the model. We adjusted the paragraph to better reflect reasons for missing predictor values.

Page 7, lines 12-22. “A model may have been derived from a prospective cohort study, whereas it will be implemented in daily practice with possibly a lower quality of data collection. Healthcare workers do not always capture the same full set of signs, symptoms and lab values, or a device that is necessary to measure one of the predictor values is unavailable. In our example, predictor values for our model were only a small part of all the information that was gathered from patients during preoperative outpatient evaluation. The PONV risk was automatically calculated by the electronic patient record during anesthesia. As the patient was anesthetized, the physician was not able to complete missing predictor values.”

5. P8 suggest that you discuss the relative weakness of a before and after design compared to the cluster randomised design, which may also account for your observed difference.
We fully agree with the reviewer that before-after designs are relatively weak and are thus more prone to observe a difference that is in fact due to time effects. However, in our situation the time periods quickly followed each other and the observed difference in behavior was very large. Moreover, the effect in behavior was risk-dependent – which would be expected – with a corresponding risk-dependent change in patient outcome. Observing an effect that is consistent with the expected mechanism of the – complex – intervention makes it much more plausible that the observed effects were due to the intervention. We added this explanation to the appropriate paragraph.

Page 9, lines 15-24. “As this was a before-after study, there is always the possibility that unobserved time effects may be the underlying cause of the observed differences. For example, of the 42 attending physicians who treated patients during the before-after study, 34 were part of the randomization of the cluster-randomized trial. As these 34 physicians received the results of the cluster-randomized trial, it is possible that this may have increased antiemetic prescription during the before-after study. Nonetheless, a large, risk-dependent change in decision making with a corresponding risk-dependent change in patient outcome makes it quite plausible that the observed difference between the before and after periods is caused by the intervention. From a study design perspective, a prediction model impact study should be regarded as a program evaluation, in which the implementation of a complex intervention is studied.”

6. Was the direction to treatment (i.e. probability threshold) in the second trial derived from patients preference, clinician recommendation, or something else?

The probability thresholds were based on existing international guidelines on PONV prophylaxis, albeit with some small adjustments according to clinician recommendations as the model that we used was different from the risk model in the guidelines (our model was tailored to the Dutch population).

Page 9, lines 9-11. “The recommendations were largely based on existing international guidelines on PONV prophylaxis, which already recommended risk-based PONV prophylaxis using a prediction model (Figure 1).”

7. P9 Is it feasible to train uses on a model's underlying assumptions or idiosyncrasies? Surely these should be either ironed out, or made obvious in the interface?

We completely agree with the reviewer. There may always be such assumptions or idiosyncrasies and ironing them out may not be that easy, as we reflect on in the last paragraph of the section “Studying patient outcome and not only effects on decision making”. Training physicians in the ins and outs of the model might be possible, but there is very little knowledge on how to do that. Hence our suggestion in the text and Box II, F.2 to provide the reasoning or research evidence behind the predicted probability. We adjusted the text to make this advice more explicit and include that there is a lack of knowledge on how to inform physicians about underlying model assumptions.
Page 11, lines 11-15. “When aware of such phenomena, one may include information on underlying assumptions of the model in the presentation and format of the prediction model. Further study is need on what the best way is to inform physicians in the underlying assumptions and mechanisms of a prediction model that is being implemented.”

8. Should the control group of interventions based around predicting prognosis be usual care, or the use of another model, or some other systematised way of thinking about prognosis (even a checklist with the question: do you think this patient is at high risk of PONV?).

We agree with reviewer that these are also very plausible alternative control groups. Which one to choose would depend on the aim of the study. We included this in the text.

Page 14, lines 18-19. “Depending on the aim of the impact study, the prediction model may also be compared to other predictive aids or interventions, rather than being compared to care-as-usual.”

9. P14: stepped wedge and cluster randomised before and after period studies are also possible

We thank the reviewer for the suggestion, which we included in the section on study design.

Page 16, lines 7-10. “The balance of a cluster-randomized trial can be improved by including crossovers in the study design, such as stepped wedge designs and cluster-randomized before-after studies, where each cluster has a time period with and without the intervention.”

10. It seems very inefficient to study each model in a new cohort. It would strike me as much more sensible to study these in situations where there is already ongoing data collection (for example hospital or nationally based audits), which have the advantage that the decisions are implicitly important (because they have been chosen as a quality metric) and the cost of data collection is borne by a non-research team (reducing cost)

We thank the reviewer for pointing that out. Making use of ongoing data collection is a very good way to reduce costs of such studies. Nonetheless, one should be aware that routinely collected data may not be optimal data for the purpose of the study – or may even be of lower quality.

Page 12-13, lines 25-27 and 1-2. “Making use of routinely collected data is also a good way to reduce costs. Process variables are often part of routinely collected data in contrast to many patient outcomes, although national registries or institutional audits may also provide the necessary patient outcomes and reduce costs. One should always be aware of possible data quality issues in routinely collected data as the data is collected for a different purpose.”
Reviewer #2: This paper discusses some problems in the implementation of an impact study. However, I feel the paper is good at exposing problems (some more or less obvious) but is not providing a clear set of recommendations / guides / suggestions on how to explore solutions.

The main message of the paper is that the effects observed from implementation studies are a consequence of a concurrency of three different kind of effects:

a) The precision of the prediction algorithm in the study population.

b) The actions that the physicians (and patients) might take after seeing the output of the model (and depending on whether these outputs are just probabilities or recommended actions).

c) The effectiveness of those actions to advert the undesirable predicted event.

The authors acknowledge that it is difficult to disentangle how much of the final overall observed effect in the implementation study is due to each of these three factors. So, for example, if an implementation of a risk score in a population does not seem to work is it because the prediction model is not accurate in that population? Or because the doctors are not reacting to the information? Or because the measures they take are ineffective? Or maybe because those measures are not effective in patients precisely scored with high risk by the model in that population?

All these reflections are important, but my worry is that the authors don’t seem to provide any suggestions on how one might try to disentangle this, so the reader is left with a sense of impossibility of carrying this kind of studies. Can we try to disentangle these effects by exploring the data in our studies? What data and associations we need to look at?

In their example they have shown this situation, but they have not tried to disentangle the mechanisms by doing further analysis in the data.

For example, did the doctors prescribed the treatment more often to patients that had certain characteristics in subcomponents of the risk score rather that looking at the overall score? (maybe to younger patients, or sicker or whatever). Was the drug less effective precisely in that kind of patients? Notice that these questions and analysis are beyond the simple original question of the trial and require a different kind of analysis. Is about dissentently the mechanisms that have operated in the decisions and effects by analysing the data.

We agree that the potential issues of impact studies, of whatever type, are listed in our paper. Unfortunately, we cannot at all and would not claim to provide solutions for all of these issues. This is also not that unexpected. The number of impact studies on prediction models is extremely limited and thus also the existing knowledge base is limited. In this paper we simply aimed to address the different challenges of model impact studies that the literature describes, that we also experienced in the two we did, and the possibilities to address these challenges – even though they are certainly not always the ultimate solutions. We changed parts of the text to be more explicit in the possibilities and suggested solutions.
On the more specific example of the difficulty to disentangle the three different kind of effects. We do not agree with the reviewer that we did not try to disentangle the different effects. We studied the first effect (predictive precision), by validating the model in our setting and updating it to be more tailored to our setting. We studied the second effect (actions and decision making) through our interviews and surveys. And we studied how changing decision making affected the results in a second study, which reflects how that changed the entire causal chain of the intervention. What we did not present in this study are our efforts to quantify the specific components and their interaction from the data that is available. When we tried to do that, we realized that we always had to make assumptions about at least one of the three components. Because of these assumptions, the analysis would not solve what was going on in the black box and its results were difficult to interpret. Through these questions, the reviewer made us aware that we have one additional suggestion: to consider additional data collection to improve the understanding of the impact study results. We added this suggestion to the text and to Box III.

Page 14, lines 4-13. “When we tried to quantify the contribution of a specific component, we always had to make one or more assumptions on how the other components affected individual patients. For example, we had to make assumptions on the prediction errors for individual patients to estimate the antiemetic effectiveness, or assumptions on individual treatment effects of (specific) antiemetic drugs were required to estimate the accuracy of the prediction model. In our example, the only way to improve our understanding of the results of our cluster-randomized trial was through either additional data collection (interviews and surveys) or through further study (the directive impact study). When designing the impact study, we would recommend considering what possible data could be collected to improve the understanding of the study’s results, especially when the results are not unequivocally positive (Box III, item 3).”

---

Finally for the specific study presented here I do not understand how exactly was designed the second randomisation: Did the same 71 doctors in the same hospitals participate? (that might explain the higher rate of prescriptions in the second experiment even in the control group).

Were the doctors randomised again or did they maintain the same groups as in the first randomisation? Did they know the results from first study? What do they mean by a "before-after" design? What were the "before" and the "after" groups?

There fact that the control group in the second study has so much higher prescriptions suggests some sort of carry-over effect from the first study.

We thank the reviewer for the suggestion to include the overlapping physicians between both studies. Furthermore, the second study was a non-randomized study, comparing clinical practice and patient outcome during the periods before and after the intervention implementation. We made the before-after design more explicit in the text and included the overlap in physicians between both studies.

Page 9, lines 5-8. “Secondly, prompted by this result, we conducted a subsequent prospective model impact study, employing a non-randomized before-after design. We implemented the
prediction model for a second time in exactly the same setting, comparing clinical practice and patient outcome during the periods before and after implementation of the intervention.”

Page 9, lines 16-19. “For example, of the 42 attending physicians who treated patients during the before-after study, 34 were part of the randomization of the cluster-randomized trial. As these 34 physicians received the results of the cluster-randomized trial, it is possible that this may have increased antiemetic prescription during the before-after study.”