Author’s response to reviews

Title: The effects of misclassification in routine health care databases on the accuracy of prognostic prediction models: A case study of the CHA2DS2-VASc score in atrial fibrillation

Authors:
Sander van Doorn (s.vandoorn@umcutrecht.nl)
Timo Brakenhoff (T.B.Brakenhoff-2@umcutrecht.nl)
Karel Moons (k.g.m.moons@umcutrecht.nl)
Frans Rutten (f.h.rutten@umcutrecht.nl)
Arno Hoes (a.w.hoes@umcutrecht.nl)
Rolf Groenwold (r.h.h.groenwold@umcutrecht.nl)
Geert-Jan Geersing (g.j.geersing@umcutrecht.nl)

Version: 1 Date: 02 Nov 2017

Author’s response to reviews:

Utrecht, November 2nd 2017

Dear dr. Siontis,

Thank you for the opportunity to revise our manuscript entitled ‘The effects of misclassification in routine health care databases on the accuracy of prognostic prediction models: A case study of the CHA2DS2-VASc score in atrial fibrillation’ and your consideration for publication in BMC Diagnostic and Prognostic Research.

We would like to thank the reviewers for reading the manuscript carefully and their many valuable comments. We strongly believe their input improved the manuscript to great extent. Most importantly, the assessment of the influence of misclassification in routine care data purely on the validation of a prediction model as the aim of the study has been clarified.

Prompted by the reviewers’ comments, we made some significant revisions to our manuscript. The analyses of HRs are now omitted, as these are not part of the traditional validation of a
prediction model and may confuse the reader. The section ‘Data analyses’ is revised, in particular the parts on calibration.

Please find a point by point reply to each reviewers’ comments below. We hope to have addressed these valuable comments satisfactorily and look forward to your reply.

With kind regards on behalf of all authors,

Sander van Doorn, MD, PhD
Karel G.M. Moons, PhD
Geert-Jan Geersing, MD, PhD
Julius Centrum for Health Sciences and Primary Care, UMC Utrecht, the Netherlands

Reviewer #1:

The aim, to examine how misclassification in electronic health records affects validation of a prediction model, is extremely important. Validation studies in routine health data are increasingly common. UK readers will be conscious that the QRISK scores replaced previous cardiovascular risk tools nationwide following successful validation in routine healthcare data.

There are three parts to this paper: (1) a study of the amount of error in the electronic health record; (2) a study of the effect of the error on hazard ratios; and (3) a study of the effect of the error on prediction rule validation studies. The first and the third are both publishable in their own right and together they make a unique and powerful paper. Including (2) as well overburdens the paper and makes it hard to follow - I would move this material to an appendix or possibly a separate paper, or remove it altogether as less novel than (3).

>>> We thank the reviewer for recognizing the novelty and power of our manuscript. Given the validation data that was collected we aimed to evaluate the influence of misclassification on the validation of a prediction rule in several ways. We agree that, although the information shown in part 2 is of interest, it takes away from the main aim of this manuscript. As such, we have gladly followed the suggestion to structure the items of our analyses (‘Data analysis’) where data on the ‘Multivariable analyses’ are omitted.

>>>

---
I struggled to follow the rationale for the analyses of hazard ratios. In the Methods and the Results they seemed to be a minor distraction in a paper that is actually about validation rather than model fitting. In the Discussion I found that the 'hazard ratio' angle was actively distracting and confusing. For example on page 14, the remarks about misclassification "averageing out" in multivariate analyses - is this a reference to the Cox models or to the validation analyses or both? While trying to decide I lost the thread.

>>> Indeed, in our manuscript we aim to focus solely on the validation of a prediction model (i.e. CHA2DS2-VASc). The remarks on page 14 referenced to the Cox models and may indeed distract and confuse. As per the comment above, we clarified our manuscript by omitting the analyses of HRs and the accompanying remark in the ‘Discussion’.

There is an established literature about how classification error affects regression models e.g. Zucker & Spiegelman (2008) Statistics in Medicine 20:1911-1933. If you are going to include the Cox model analyses in this paper then it seems an omission not to discuss the results - in the Discussion section - in terms of this established literature.

>>> As mentioned above, we omitted the analyses of HRs from Cox models from the revised manuscript to not further confuse the reader. As such, methods for correcting for misclassification in Cox modelling is not discussed in the Discussion.

However, the publication as mentioned by the reviewer provides very valuable information on this topic. We now mention it in the ‘Discussion’ as a remark for readers interested in Cox model development with potentially misclassified predictors:

** Last, it should be stressed that we only focus on the validation of a prediction model. For prediction model development using Cox analysis, methods on how to correct for misclassification in predictors have been previously addressed.[29]

In the Introduction, you cite several references 3-7 for measurement error models. I have usually seen the terminology of 'measurement error' used for the case of continuous variables and the term 'classification error' for categorical or binary variables. Your reference 5 is the Carroll book - I only have the first edition to hand, which includes less than a page on classification error, but perhaps the second edition has more? Likewise the book by Wayne Fuller (which your software has accidentally cited as Wayne AF instead of WA Fuller!) emphasizes continuous variables, doesn't it? I don't have my copy of Fuller to hand so accept my apologies if I'm wrong about that; if I'm right, could you reference specifically the chapters or sections that are most relevant to dichotomous variables?
Thank the reviewer for this careful reading. We have indeed added references that mostly address the effects of measurement error on studied associations. Prompted by this excellent comment, we have now removed some of the original references and added others:

• Buonaccorsi JP. Measurement Error. CRC Press; 2010.


The first paragraph on page 14 ends with a remark that "validation studies ... commonly report the observed risk per score" and that at score 1 "this risk was nearly twice as high" as the reference value. This doesn't seem to match with Table 4, and in fact Table 4 suggests that numbers of events are too small, at the lower score levels, to make any comment on observed risk.

We thank the reviewer for pointing this out. We regret the first column of Table 1 indeed contained an error, where the rows should have read score 0-9. We now present the correct scores, showing that the incidence rate (IR) is 0.4 for CHA2DS2-VASc score 1 based on the index predictors and 0.7 (nearly twice as high) for the reference predictors. We agree that numbers are small, which we also extra acknowledge in the new Discussion.

When calculating c-statistics for the model, did you use a c-index for censored data or did you assume follow-up was complete to two years?

For calculating discriminative ability of the CHA2DS2-VASc model based on the index or reference predictors, under the section newly entitled ‘Discrimination’, we used indeed the c-index for censored data. This is now clarified in the ‘Data analyses’ section.

You might attract criticism for using mortality, rather than stroke, as the outcome in a CHADS2 project, but I agree with you that for this methodological study of misclassification error validation against mortality can still be informative. There are already clinical papers assessing
the usefulness of CHADS2 for predicting mortality (e.g. Kurtul & Acikgoz 2017 A J Cardiology 120:8-14 or Yang et al 2016 PLOS One) albeit in more acutely ill patients.

Indeed, we chose this endpoint purely for methodological reasons. To prevent the potential criticism and prompted by this comment we now state this more explicitly at various places. In the ‘Abstract’ section:

** AIM: To quantify the amount of misclassification in routine care data and its effect on the validation of the existing risk prediction model. As an illustrative example, we validated the CHA2DS2-VASc prediction rule for predicting mortality in patients with atrial fibrillation (AF).**

In the ‘Methods’ section:

** Our aim was to study potential misclassification in the prediction variables, not in the outcome. The CHA2DS2-VASc was originally developed to predict either ischaemic stroke, peripheral embolism, or pulmonary embolism. In our methodological study, however, we used all-cause mortality as an illustrational outcome for two reasons. First, stroke may be difficult to diagnose, especially stroke as the cause of (unexpected) death. The outcome all-cause mortality can be objectively determined. Second, such mortality data may often be captured by the municipal authorities, as was the case for the general practices in our study, further avoiding misclassification in the outcome. We manually checked vital status using the electronic patient file. Follow-up was a minimum of two years.

Given that the main aim of the paper (according to the abstract) is about validation, rather than model fitting, it seems disproportionate that 2 of the 5 items under "Data analysis" are for model fitting. I would strongly recommend you give more detail on the validation methods. Firstly, on page 9 item 4 under "Data analyses" says "We then used Cox proportional hazards models to analyze the predictive performance of both scores." Many readers may be aware of the Cox modelling for deriving a model but not aware of its use in calibration so you may need to give slightly more detail and a reference e.g. chapter 15 of the Steyerberg book.

We thank the reviewer again for recognizing the main aim of our study, i.e. validation of a prediction model rather than model development, and for the suggestions. Indeed, most readers will be aware of Cox modelling for deriving a model and more detail on its use in our study may therefore be needed. We now present more detail on the validation methods, most importantly for assessing calibration. This was not possible for the original CHA2DS2-VASc model, as no data on the baseline hazard or model coefficients are available for CHA2DS2-VASc predicting mortality. For this reason we first fitted a CHA2DS2-VASc model ourselves using the reference predictors. Obtaining the baseline hazard and hazard ratios from this fitted model, we then
validated the model using the index predictors. We agree that by describing the process of fitting
the model CHA2DS2-VASc model may cause confusion but it was only done to make the
validation of the score in our data possible. We now clarified the relevant section under ‘Data
analyses’ and the terminology under ‘Results’ (i.e. refrained from using the term ‘develop’):

** 4. For assessing the influence of misclassification on calibration, data on the baseline hazard
and hazard ratios for the CHA2DS2-VASc model predicting mortality are missing. To obtain
these, we first fitted a multivariable Cox proportional hazards CHA2DS2-VASc model using the
individual reference predictors. We assessed calibration by creating a calibration plot and
calculating the calibration slope. Using the same baseline hazard and hazard ratios, we than
assessed calibration using the index predictors; the difference in calibration then occurring can
only be caused by misclassification.

Likewise, there are surprisingly few validation metrics in the results (have I missed some?).
There are numerical values for c-statistics, in the text, the concordance table in Figure 1, and a
graphical representation of calibration in Figure 2. It would greatly enhance the applicability of
the paper to include the P/O statistic for overall calibration, and the currently fashionable
measures calibration slope and calibration intercept. Although I agree with Royston & Altman
(2013, BMC Med Res Methodology, 13:33) that the former is actually a measure of
discrimination or fit rather than calibration, it is informative and increasingly widely used even if
it is not well-named.

>>> Indeed, the ‘traditional’ validation of a prediction model should include a measure of
discrimination (e.g. c-statistic) and a measure of calibration (calibration plot, with intercept
and/or slope; test statistics).[1, 2] In addition to the calibration plot, we now also report
calibration slope.

Perhaps consider too whether you wish to report on sensitivity & specificity (or PPV) for the
CHADS2 score at a specific threshold.

>>> We appreciate this suggestion, as very often a certain threshold is used for treatment
decisions in clinical practice. However, specifically for CHA2DS2-VASc the most appropriate
threshold for initiating or withholding treatment (i.e. anticoagulants to prevent stroke) is still
under heavy debate[3-5] partly due to heterogeneity in existing validation studies.[6] To refrain
from engaging in this ongoing discussion, we propose to present the results for the full
CHA2DS2-VASc score only.
Although my comments above are extensive, this is an excellent piece of research. The depth of your work on the data and the decision to address validation make a unique and important manuscript.

>>> We thank the reviewer for his/her positive remarks and also on the valuable comments to further improve and focus the aims of our work.

Reviewer #2: This paper considers the sensitivity of the CHA2DS2-VASc score to errors in the predictor variables. Specifically, the paper performs a validation of this score using both gold standard (reference) and potentially misclassified (index) predictors.

I am not convinced that the CHA2DS2-VASc score is the best example to use for this case study. For one, the score does not seem to actually make predictions (Authors: if it does, please add this to the Methods section and Table 1). Instead, it assigns patients to an integer score between 0 and 9 which, presumably, can be used to rank patients in terms of severity / expected mortality. Second, it is a relatively simple score with only 7 (binary/ordinal) predictors with simple integer weightings. Of these, misclassification is only considered in 5 of the predictors. Since the weightings are so simple (either 1 or 2), misclassification in a couple of the predictors may actually cancel each other out.

>>> Indeed, the CHA2DS2-VASc model assigns patients with a score 0-9 upon which a subsequent treatment is advocated (no treatment for a low score, anticoagulants for a high score). However, the underlying principle certainly is based on absolute predictions. In addition to the original development paper,[7] many existing validation studies of CHA2DS2-VASc were done on data obtained from routine care, and they commonly present expected risks (i.e. incidence rates) per score, and the decision to initiate anticoagulant treatment is based on an absolute risk threshold. Moreover, the CHA2DS2-VASc model is widely advocated in practice guidelines and indeed used in daily practice over the world. Although we fully understand the concern of the reviewer, for these reasons we think that the chosen example is interesting to make our points. Prompted by this comment we do now include explicit information on the expected risks in the Table 1. and ‘Methods’. Also the reason why we chose for the CHA2DS2-VASc model is stated under ‘Strengths and limitations’:

** Strengths of our study include the opportunity to assess misclassification in predictors from routine healthcare from the well-known CHA2DS2-VASc model. This model is recommended by multiple practice guidelines, and frequently validated using data from routine healthcare.
Furthermore, we do agree that the integer weighting of predictors in CHA2DS2-VASc is (over)simplistic, and as such could easily cancel out misclassification. However, even for a simple model such as CHA2DS2-VASc, we observe substantial differences in the score as assigned by index or reference predictors (Fig. 1). We now discuss this in further detail in the ‘Discussion’:

** Third, the CHA2DS2-VASc score is a simplistic decision rule, with limited integer weighting of predictors (1 of 2 points). Although we did find pronounced differences in the score as calculated with index or reference predictors, using such simple weighting could also have ‘cancelled out’ some of the misclassification. Future studies should investigate the effects of misclassification in predictors on the predictive performance of other prediction models.

And:

** Future studies should focus on the influence of misclassification on the predictive performance of more complex models, or the influence of different predictor misclassification patterns, e.g. using a simulation study.

I have a number of concerns regarding the analysis section and the motivation behind each analysis. I will discuss these by table and figure.

Table 3: I am not convinced that the various measures presented in this table (i.e. kappa to NPV) are actually that helpful for summarizing the misclassification. I think it would be far clearer if the actual data (e.g. the information from the 2x2 tables) were presented in this table. This should be straightforward as there are only 5 predictors. If the diagnostic measures (sensitivity to NPV) are to be retained, it would be helpful if more interpretation was provided.

>>> We thank the reviewer for these suggestions. Indeed, we could provide the raw data, but this would certainly lead to larger tables, and perhaps somewhat unnecessary data. We think that the measures presented in Table 3 (sensitivity, specificity and predictive values) are fairly easy to interpret and indeed are the relevant summary measures to report. However, to further assist the reader, prompted by this comment, we provide now additional interpretation on these measures in the ‘Results’ section. Furthermore, we now present the 2x2 tables with the actual data (as suggested by the reviewer) in Supplement 1.

** For cross tables with the presence and absence of each predictor individually, see Supplement 1. Sensitivity (i.e. the proportion of patients with heart failure according to the reference predictor that correctly had the diagnosis according to the index predictor) was lowest for heart failure (55%) and highest for diabetes (89%). Specificity (i.e. the proportion of patients without heart failure according to the reference predictor that correctly were diagnosed as such using the
index predictor) ranged from 83% (hypertension) to 99% (diabetes). A similar pattern was observed for the predicted probabilities. Diabetes showed the highest PPV (i.e. the probability of having diabetes according to the reference predictor if diagnosed with diabetes according to the index predictor) and NPV (i.e. the probability of not having diabetes according to the reference predictor if the index predictor was absent) of 98.8% and 96.4%, respectively. Hypertension again showed the lowest values (83.3% and 81.6%, respectively).

Table 4: It is not clear whether this analysis and table are necessary. The results are not discussed in the text.

>>> Table 4 shows the number of person-years, the number of events and the observed incidence rate (IR) for each CHA2DS2-VASc score. Although these data are indeed not part of a traditional prediction model validation study,[1] many studies validating CHA2DS2-VASc model (including using routine care data) do present these incidence rates (see comment above).

We therefore think that these data may be of importance, especially to the clinical audience. But prompted by this comment, we now discuss this table in more detail under ‘Results’:

** Table 4 shows the number of patients, the number of events, the total number of person-years and the observed IR of all-cause mortality for each CHA2DS2-VASc score calculated with index and reference predictors. Although small numbers for the lowest and highest score limit definite conclusions, we observed a relative ~10% difference between both sets of predictors. For instance, for patients with a score 4 according to the index predictors the IR was 5.6 per 100 person-years, while this was 6.5 per 100 person-years for the same score according to the reference predictors.

Table 5: I do not understand why new Cox models are fitted to the reference and index predictors since this is (I presume) a validation exercise rather than an exercise in model development / updating. The analyses reported in this Table certainly need to be motivated better. Additionally, I also don't really understand why age and sex are intentionally left out of the re-fitted models.

>>> We regret that the analyses of HRs are unclear, and distract from the study as a validation exercise. As mentioned above, also elaborated upon in our answers to reviewer #1, we made significant revisions to the ‘Methods’ and ‘Results’ sections:

- The analyses of HRs are omitted as they are not part of a traditional validation study of a prediction model.
• The methods used for assessing the influence of misclassification on calibration are rephrased, see above.

• Details on the model fitted to assess the influence of misclassification on calibration are specified in Supplement 2.

• The subheading under ‘Results’ now read ‘Discrimination’ and ‘Calibration’.

Figure 1: It might be better to provide both numbers and row percentages (please state that these are row percentages in the caption). Also, more insight could be provided in the text, e.g. how many misclassifications cancel out?

>>> We thank the reviewer for this comment. We now present both numbers and percentages in Figure 1, and state so in the caption.

Figure 2: Please see my comments for Table 5. This perhaps shows that the CHA2DS2-VASc score is not suitable for this case study as it does not produce mortality predictions and hence cannot be assessed with respect to calibration.

>>> Indeed, although the CHA2DS2-VASc score was originally not indented to predict mortality this outcome was chosen for methodological reasons, as mentioned in the manuscript. We believe that, in the absence of data on the baseline hazard and model coefficients for CHA2DS2-VASc predicting mortality, the method employed in our study does illustrate the difference in calibration due to misclassification in the predictors, since we use the same baseline hazard and hazard ratios to assess calibration using index and reference predictors, respectively. However, following this comment, the terminology in the original manuscript has been revised for clarity, see above.

There is also, I think, an omission in the Results section. I couldn't find estimates of the c-index for the CHA2DS2-VASc score based on the index and reference scores respectively. To be clear, since this is a validation exercise, models do not need to be fitted to produce these values.

>>> The c-statistic for CHA2DS2-VASc was indeed somewhat placed ambiguous in the original manuscript. We now present the c-statistic for CHA2DS2-VASc as a (simplistic) point-based score in a section entitled ‘Discrimination’.
Some other more minor points:

Abstract | Aim: As mentioned earlier, the CHA2DS2-VASc score does not seem to predict (absolute) mortality.

>>> We now clarified this in the abstract, see our comments above.

Abstract | Conclusion: This conclusion is far too strong and almost certainly does not generalize.

>>> We thank the reviewer for this suggestion. We rephrased the conclusion of the ‘Abstract’ as follows:

In a case study validating the CHA2DS2-VASc prediction model, we found substantial predictor misclassification in routine healthcare data with only limited effect on overall model performance. Our study should be repeated for other often applied prediction models to further evaluate the usefulness of routinely available health care data for validating prognostic models in the presence of predictor misclassification.

Background (l7): Please add references.

>>> We added the appropriate references.

Background (l44): Please add a reference for the model.

>>> We added the appropriate reference.

References


