Reviewer’s report

Title: Development and validation of a prediction model to estimate the risk of liver cirrhosis in primary care patients with abnormal liver blood test results: protocol for an electronic health record study in Clinical Practice Research Datalink

Version: 0 Date: 25 Feb 2019

Reviewer: Maarten van Smeden

Reviewer’s report:

The authors should be commended for their initiative to write a detailed protocol for the development and validation of a prognostic prediction model for liver cirrhosis. I have a couple of remarks, particularly regarding the planned validation. My remarks are found below.

The protocol states that (L210) "We will omit 30 GP practices from our model development dataset and use these to externally validate our prediction model" and "The omitted practices will be chosen prior to model development." I have two remarks on this.

a. It is my view that the 30 to be omitted practices should either: already have been identified and listed in this protocol, or, the exact criteria for partitioning should be defined in the protocol.

b. It is unclear to me why the authors opt for the split sample approach instead of a cross-validation or bootstrap approach to approximate out-of-sample predictive performance (related to my second point below)

The "validation" of the model seems way too ambitious for the data at hand (~250 expected events in the test set, approximately 8.3 events per center). The authors state: "we will evaluate the model performance in each region and GP practice." With 8.3 events per center there will be not much room for validation. Also, the validation in subgroups of e.g. age (subgroups not defined, but should be) and BMI will render much lower events than the recommended number (200, Collins et al., Statistics in Medicine, 2015, 10.1002/sim.6787). I encourage the authors to the following approach: consider developing the model by including predictors that may explain potential variation in predictive performance in the modeling and perform a cross-validation or bootstrap approach to evaluate the performance in the whole set and potentially in smaller groups (avoid split sample validation).

I have difficulty understanding the rational for using the heuristic shrinkage factor combined with backward elimination. Can the authors explain why they don't opt for a penalized regression approach (e.g. lasso or elastic net) that incorporates the shrinkage in the estimation of the model?

L427: please avoid showing the calibration slope in the development set. This can only mislead the audience.
Level of interest
Please indicate how interesting you found the manuscript:

An article of importance in its field

Quality of written English
Please indicate the quality of language in the manuscript:

Acceptable

Declaration of competing interests
Please complete a declaration of competing interests, considering the following questions:

1. Have you in the past five years received reimbursements, fees, funding, or salary from an organisation that may in any way gain or lose financially from the publication of this manuscript, either now or in the future?

2. Do you hold any stocks or shares in an organisation that may in any way gain or lose financially from the publication of this manuscript, either now or in the future?

3. Do you hold or are you currently applying for any patents relating to the content of the manuscript?

4. Have you received reimbursements, fees, funding, or salary from an organization that holds or has applied for patents relating to the content of the manuscript?

5. Do you have any other financial competing interests?

6. Do you have any non-financial competing interests in relation to this paper?

If you can answer no to all of the above, write 'I declare that I have no competing interests' below. If your reply is yes to any, please give details below.

I declare that I have no competing interests

I agree to the open peer review policy of the journal. I understand that my name will be included on my report to the authors and, if the manuscript is accepted for publication, my named report including any attachments I upload will be posted on the website along with the authors' responses. I agree for my report to be made available under an Open Access Creative Commons CC-BY license (http://creativecommons.org/licenses/by/4.0/). I understand that any comments which I do not wish to be included in my named report can be included as confidential comments to the editors, which will not be published.

I agree to the open peer review policy of the journal