Reviewer’s report

Title: Predictive Modeling of Disease Burden in Adult-Onset Ulcerative Colitis Using Health Administrative Data

Version: 0 Date: 25 Nov 2018

Reviewer: Glen Doherty

Reviewer's report:

The authors present a study of health administrative data from adults patients with newly diagnosed UC from a single Canadian centre. The data were analysed with a view to creating a model predictive of disease burden (either extensive and/or severe UC). The results are interesting but even the best models performed only fairly in the discrimination of patients with high disease burden. The authors take what could be considered an excessively optimistic slant in their conclusions, given that they do not present any validation data which is the single biggest weakness. Rather than suggesting (as they do in the conclusion of the abstract) that these models could be applied in future population based studies I think the conclusion should rather be that they need to be applied in other cohorts for validation. I'm concerned that in the records review of the Ottawa patient cohort it is not stated how many patients had a prospectively documented endoscopic Mayo score. If the severity was retrospectively inferred in a high proportion this will undermine confidence in the model. The other big issue, which is touched on but not adequately discussed, is the fact the disease burden of the cohort does not reflect that seen in population based studies (only 20% with proctitis, almost 50% with high disease burden); there is huge potential for bias arising and again this underlines the need for the model to undergo validation in a representative population based cohort before it can be applied more widely. A few other minor issues; there are a number of errors where words are duplicated in the text (e.g. Page 5, linkagelinkage, datadata). On page seven the authors describe that total disease burden is 'a combination of extent and activity' but there is no precise description of how the two metrics are combined. On page 9, I find the definition of 'high disease burden' confusing; can I suggest for clarity listing each category that is included (e.g. moderate extensive, severe extensive and severe left sided; correct?). Overall the manuscript is otherwise well written but I think the Discussion could be shortened.

Are the methods appropriate and well described?
If not, please specify what is required in your comments to the authors.

Yes

Does the work include the necessary controls?
If not, please specify which controls are required in your comments to the authors.

Unable to assess

Are the conclusions drawn adequately supported by the data shown?
If not, please explain in your comments to the authors.

Yes

**Are you able to assess any statistics in the manuscript or would you recommend an additional statistical review?**

If an additional statistical review is recommended, please specify what aspects require further assessment in your comments to the editors.

I recommend additional statistical review

**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 have no conflict of interest relevant to this study to declare

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