Reviewer's report

Title: Estimation of cumulative number of post-treatment Lyme disease cases in the US, 2016 and 2020

Version: 0 Date: 01 Oct 2018

Reviewer: Gerrit Gort

Reviewer's report:

The authors wrote an interesting paper about the estimation (as I would say) of the cumulative number of PTLD cases in the US, which to me appears to be relevant for public health. It is well written and concise. Of course I have a number of remarks, firstly some more important remarks and secondly some smaller remarks.

The outcomes of your simulations depend heavily on the 329,000 new cases per year, which are based on a health insurance claims database (P4). Are you sure that these 329,000 are really new cases, and not duplicate diagnostic tests (e.g. for second opinions) or replicate tests for earlier diagnosed cases? If that would be the case, the number of actual new cases may be substantially lower. In other words, how realiable is this rate?

To continue with the 329,000: in your simulation you assume Poisson variation. Poisson variation is, relatively speaking, very small for such large numbers: a Poisson distributed variable with mean 329,000 has variance 329,000, so standard deviation sqrt(329,000)~566. Therefore, you will sample counts with a 2SD range of 328,000 +/- 2x566 ~327,000-329,000. That is not expressing a lot of uncertainty about this number. Would a larger variance not be much more realistic, if it is to express uncertainty about the number of yearly new cases of LD?

The authors work with two separate treatment failure rates (10% and 20%; P5 top), allowing deviations from these numbers in the (separate) simulations, based on beta distributions. But the numbers 10% and 20% indicate that is a lot uncertainty about the failure rate. Why do you take this approach, and not work with a single, but highly uncertain failure rate?

The authors try three different scenarios (Pp 5-6). Scenario A is VERY unrealistic: a gradual increase from 0 to 30,000, and then suddenly an enormous jump to 329,000 new patients per year. Why do you include such an unrealistic scenario?

The authors assume that survival rates for patients with PTLD are the same as for the general populations (P6 bottom). How realistic is this? Can you supply a reference for this assumption? May be PTLD patients experience high mortality due to other causes which become lethal in their situation.
The deterministic estimate of prevalence (P8) is not "validating our results" (P4) in a general sense, but only as a check for the simulations. I think the wording on page 4 is too strong.

In the discussion (P11, bottom) you mention that the expected number of new LD cases is critical.

An alternative approach to use could lie in Bayesian statistics (although I cannot see the exact workings for it directly here), as it incorporates all sorts of uncertainty to arrive at a posterior distribution for the quantity of interest. May be you can make some reference to this in the discussion.

Smaller remarks:

P2 L22: (and others) You use the word "calculated" in many places. For me a better word is "estimated": you obtain an estimate for quantities (namely, the Nth year prevalence of PTLD) with related uncertainty.

P2 L22-24: the cumulative numbers of [ ] the prevalence of?

P3 L43-47: You mention here that the "precise societal burden …has never been adequately quantified". And a bit further: "To address this absence…". This suggests that your estimation is giving a quantification of the precise societal burden. Do you really think that this is the case? Or are you just adding a small contribution?

P3 L51: you mention here "mathematical modeling". I see more statistical modeling, as you use probability distributions (like binomial, multinomial, uniform and beta) to quantify uncertainty.

P4 L6: in what sense are the inputs "conservative"? On the low side or on the high side?

P4 L9: please specify why you think that your statistical assumptions are "appropriate"; sometimes I have some doubts (e.g. with Poisson distribution, see earlier)

P5 L22: it may be good mention here already that you assume that death rates for PTLD patients are identical to these from the general population.

P5 L 33: mention that this is about incidence of LD (and not PTLD)
P5 L35: the word "arguably" is essential here, because scenario A is completely unrealistic (see above).

P5 L53: would exponential growth ever be a realistic option in the expanding phase of a disease like LD, which is non-contagious? If not, it doesn't make much sense to make the comparison here, suggesting that linear growth is highly conservative.

P6 L40-49: The part about the use of the beta distribution is a bit unclear. May be you can give an example, to give the reader in impression of the uncertainty you introduce here (e.g. with $p=0.10$, you sample a $p$ from beta(50,450) which has mean 0.10 and SD 0.013, so roughly 2 SD range for $p$ is 0.074-0.128).

P9 L11: explain how you calculate a "CI". I guess CI stand for confidence interval, but note that confidence intervals in statistics have a special meaning. Here you produce the 2.5 and 97.5 percentiles from the 500 simulated results, aren't you?

P9 L40-42: Figure 1 is not showing that "the relative distribution by age and gender was insensitive to failure rate", because you show the lumped results from the simulations for 10% and 20% (I guess).

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

Unable to assess

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.

Not relevant to this manuscript
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