Reviewer's report

Title: An extensible framework and database of infectious disease for biosurveillance

Version: 0 Date: 15 Mar 2017

Reviewer: David M. Hartley

Reviewer's report:

This study addresses an important and difficult topic, namely the systematic classification of a wide range of diseases of humans, plants, and animals for the purposes of biosurveillance. Such a classification should enable efficient and accurate monitoring, tracking, and alerting of diseases of interest to a wide variety of users of biosurveillance. To date, this problem has not been solved, so any progress in the area is of interest to researchers. The material described in this paper is, therefore, important to disseminate to the readership of BMC Infectious Diseases.

However, the manuscript is not presently suitable for publication as presented. The paper is difficult to follow as currently organized, and several passages seem limited or incomplete. It is unclear whether the authors executed a comprehensive, valid gaps and requirements process. If so, then describing that process early in the study would be helpful. There is an emphasis on vector borne disease, but the reason for this is not identified in the study. The paper might be clearer if restructured so that 2-3 (or more) illnesses are coded according to the proposed classification scheme, perhaps a vectorborne illness (say dengue, malaria, or Rift Valley fever), a respiratory illness (say influenza or measles), and a waterborne infection (say cholera or crypto) to illustrate more clearly; that might remove the ambiguity identified in the comments below. The incorporation of use case scenarios would also help the reader to understand the rationale for including different facets of the proposed classification scheme.

Specific comments

Page 1 line 45: "... but biosurveillance often falls under global health security". The meaning of this phrase isn't clear. Biosurveillance is relevant to global health security, as numerous publications relating to biosurveillance studies in the international domain demonstrate. This sentence might be rephrased to read something akin to "... and studies illustrate the relevance of biosurveillance to both national and global health security."

Page 2, first full paragraph: The upshot of the end of the paragraph is that the ontologies of infectious disease discussed are more applicable at the micro or intra-host level as opposed to the macro or intra-host level. Is the purpose of this discussion to identify gaps in existing ontologies, which are then to be filled in the research described in the next section? How do these issues
relate to the BRD? In the following paragraph, the term "classification system" is used, but it is unclear if this system is a new ontology, a revision to existing ontologies, or a different scheme altogether.

Page 2, "Domain requirements": How were these requirements identified? Are these acknowledged by researchers in the field as found by survey or questionnaire? Are they documented in other studies? Were they inferred from a critical literature review? What was the requirements definition process? The reader is left wondering not only where these come from, but whether they are complete — are there other requirements that are equally (or more) important?

Page 2, first bullet: These are important points. In fact, the issue of synonyms becomes harder when multiple languages are considered. Several biosurveillance systems collect material in different languages at present, but no discussion of the language issue appears in this passage.

Page 3, first bullet: The discussion concentrates on mosquitoes, which is too limiting. Avian pox, for example is spread via mechanical transmission through biting insects generally, including mosquitoes of multiple genera. More importantly, what about indirectly transmitted diseases where there are no vectors, such as waterborne (e.g., cholera), blood borne, food borne, etc? Perhaps the scheme should have a slot for method of transmission more generically, with increasing specificity possible at different levels within this slot?

Page 3, third bullet: The authors state that "Much of biosurveillance takes place under an umbrella of syndromic surveillance" and undoubtedly this is true, but throughout the paper it seems that biosurveillance based on the mass collection of open source news media and related materials from the Internet is neglected. How do these systems, which have been reviewed in the literature, relate to the work in this study? Also, it's not clear what the phrase "represent syndromes in the same fashion as diseases" means.

Page 4, vector bullet: Why is this separate from "Transmission", which is discussed in a following bullet? Overall, it seems like vectorborne infections are stressed in this classification relative to respiratory, waterborne, etc illnesses. If true, why is this?

Page 5, disease parent bullet: How is malaria a respiratory disease?
Page 5, first complete paragraph: The term "organism table" appears suddenly. The example that follows again focuses on mosquitoes. Is there a different organism table for agent, vectors (when present) and hosts? Note also that it is unclear that carrier hosts, e.g., environmental reservoir hosts that may or may not become symptomatic, are included in the proposed classification scheme. Where would they appear — under population?

Page 5, next paragraph: The phrase "Just as organisms have self-referencing ties allowing a hierarchy," is jargon and not likely to be understood by readers of BMC Infectious Diseases. What does this mean? Continuing in the same paragraph, for the case of influenza, does the hierarchy follow a virological structure, or a host structure? Influenza typically is classified according to type (A, B, C) and for the case of A according to H and N surface proteins, etc. The example given, however, seems to stress the host role (*avian* influenza). How does this relate to recognized/traditional schemes?

Page 5, next paragraph: Is there a distinction between "tags" and "flags"?

Page 5, last full paragraph: What is the Django application?

Page 6, first paragraph: Probably no readers will dispute that one is able to link conditions to ICD codes, but what is the rationale for doing so? What is the use case? Is it to include ICD codes from actual clinical records (e.g., EHR), or to exploit ontology structure and and associate ICD codes to particular illnesses? Or both, or neither?

Page 6: The first two paragraphs seem more like Results than Discussion.

Page 6, bottom: The section "Future Directions" seems repeated from the above section, "Discussion and Future Directions".

In the figure captions, "Malaria" and "Anthrax" should not be capitalized.

The references need to be formatted correctly; as of now they contain BibTeX codes.
Are the methods appropriate and well described?
If not, please specify what is required in your comments to the authors.

No

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.

Unable to assess

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:

Needs some language corrections before being published

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