Author's response to reviews

Title: PHSkb: A Knowledgebase to Support Notifiable Disease Surveillance

Authors:

Timothy J Doyle (tdoyle@cdc.gov)
Haobo Ma (hma@cdc.gov)
Samuel L Groseclose (sgroseclose@cdc.gov)
Richard S Hopkins (Richard_Hopkins@doh.state.fl.us)

Version: 2 Date: 19 April 2005

Author's response to reviews: see over
April 19, 2005

Timothy Doyle
Centers for Disease Control and Prevention (CDC)
1600 Clifton Road, MS E-91
Atlanta, GA 30333
tdoyle@cdc.gov

Editor
BMC Medical Informatics and Decision Making
editorial@biomedcentral.com

Re: MS: 4551797358246130 – PHSkb: A Knowledgebase to Support Notifiable Disease Surveillance

Dear Editor;

My coauthors and I would like to thank you for reviewing the above referenced manuscript. We found the reviewers’ comments helpful in improving the overall quality of the manuscript. Please consider the revised manuscript for publication in BMC Medical Informatics and Decision Making.

We would first like to clarify a logistical matter that partially addresses some of the reviewers’ comments. Our overall aim is to publish the PHSkb database and to make it as widely available as possible. However, we wanted to coordinate the release of the PHSkb with a thorough, yet concise description of the methods used to construct it. The manuscript, formatted as a database article, is not an end in itself. Rather, it is intended primarily as supporting documentation for the benefit of those accessing the knowledgebase. Our hope was to have the public release of the PHSkb coincide with publication of the manuscript describing it. In our initial letter to BMC editors, we indicated that the PHSkb could be made available to the reviewers if necessary. Since some of the reviewers’ comments suggest an incomplete understanding of the structure and content of the PHSkb, or the view that the manuscript is an end in itself independent of the database, we have now posted the database files to the FTP site for access by the reviewers. The files can be opened in Protégé software, which is available as a free download from the Protégé website http://protege.stanford.edu/index.html. We regret any inconvenience you or the reviewers had in trying to locate the PHSkb files at the FTP site, or in interpreting the manuscript content without access to the knowledgebase itself. The original reviewers we suggested for the paper included several people with knowledge of Protégé software. We further hope the present reviewers do not find it burdensome to access the knowledgebase files using Protégé.

A point-by-point reply to the reviewers’ comments follows below with the reviewers’ comments in bold text and our reply in plain text.
Referee 1

This paper provides an interesting model for notifiable diseases, however it has not demonstrated that PHSkb will really contribute to improve sensitivity, timeliness and quality of notifiable disease surveillance as it is initially presented. This should be more clearly stated.

The reviewer is correct in noting that the paper itself does not demonstrate improved surveillance sensitivity, timeliness, or quality as a result of this knowledgebase. Rather, these anticipated outcomes were the original rationale for creating the knowledgebase. We have tried to state this more clearly by inserting additional text into the manuscript.

In the background section, final paragraph, we have reworded the third sentence to read, “The long-term goal of the…” Also in the background section, end of final paragraph, we have added the following sentences:

“This paper describes the initial creation of the PHSkb. Further field testing will be needed, however, to determine the impact of such methods on surveillance sensitivity, timeliness, and data quality.”

In the discussion section, final paragraph, after the first sentence, we have added the following sentence:

“Field testing of the PHSkb is needed, however, to determine its impact on surveillance metrics such as sensitivity, timeliness, and data quality.”

Also in the discussion section, we have added the following sentence to the end of the last paragraph:

“Revisions to the content and structure of the PHSkb should be guided by user feedback and the results of field testing.”

To date, content of PHSkb is not validated with state surveillance partners. There may be a long way before this tool is used on an ongoing basis, since it is still not distributed in the USA jurisdictions.

We acknowledge that it may be some length of time before the tool is used on an ongoing basis. However, a significant step in distributing the PHSkb and making a broad range of partners aware of its existence is publishing an accurate description of the methods used to create the knowledgebase in the peer-reviewed literature. No changes were made to the manuscript as a result of this comment.

Referee 2
This paper addresses different problems… I think that the paper suffers from this disparity of questions, and should be centered more on the use of an ontology to model the disease surveillance.

Our main objective is to publish the PHSkb database with enough descriptive information contained in the manuscript to allow readers to understand the basic structure, content, and rationale for the database. The paper was originally formatted and submitted as a database article type. We followed the required format for database article types specified by BMC. This format includes sections for background, construction and content, utility, and discussion. Within these sections, authors are further instructed to include information on, “the perceived need for the database…stating its potential value…information on data sources…intended uses of the database…planned future development…importance and relevance of the database.” We attempted to adhere to these instructions by including relevant information in the manuscript.

We empathize with the reviewer’s comments. Database articles are a bit like articles on music criticism. The reader is separated from the substance of what is being described. Ultimately, such articles are a poor substitute for listening to the music oneself, or in this case, seeing the database itself. The PHSkb database itself is the main focus of our attention. We hope that once the reviewers have the opportunity to see the database via the FTP site, they will recognize that it has always been the focus of the manuscript describing it. No changes were made to the manuscript as a result of this comment.

For example, Tables 1 to 6 are descriptive information that one would expect to find on some official public health Website rather than in a scientific journal (I was surprised : Is Table 2 ‘original’ ? it looks as a very basic description of the state surveillance that should be on the CDC website ?).

Tables 1 to 6 are all original. They did not exist prior to this project and have not been disseminated previously. Since data in the PHSkb still needs to be validated with state surveillance partners, we do not foresee posting the information from tables 2-6 to the CDC website, outside the context of this manuscript and accompanying knowledgebase.

Table 1 provides links to the source information used to instantiate the state specific reporting requirements in the PHSkb. This allows the reader to verify for oneself the original state specific data in its original format. Tables 2 thru 6 are artifacts produced from the knowledgebase. Such information did not previously exist in this format, prior to creation of the knowledgebase. If the reader wanted to obtain the information contained in tables 2-6, they would have to either query the knowledgebase, or do as described in the manuscript and obtain each states reportable disease requirements and harmonize term variants to a standard. Tables 2-6 are a poor substitute for the richer knowledge representation contained in the PHSkb. However, the advantage of these tables is that they allow the reader to understand the breadth of the notifiable disease domain addressed by the knowledgebase, without the need to download the Protégé software and understand how to use it. This is an important consideration for a large
proportion of readers who may come from a public health surveillance background and may have limited informatics skills.

We agree that Tables 1 to 6 are not profound from a scientific perspective. Once again, however, we wish to call the reviewers attention to the fact that the manuscript was submitted as a ‘database’ article type. Under the traditional circumstances of a hard-copy print journal format, we would not consider submitting such lengthy tables for publication. However, we deliberately chose *BMC Medical Informatics and Decision Making* for our database article because, as an all-electronic journal, such lengthy tables can be easily uploaded as additional files for interested readers. This enhances the descriptive capability of the inherently challenging ‘database’ article type. Tables 1 thru 6 are not meant to appear in the body of the text, but rather to be available as additional supporting files. We apologize for any confusion regarding this distinction. We have made changes throughout the text to refer to these as additional file 1 thru 6 instead of table 1 thru 6. In addition, we have renumbered table 7 to table 1, and included it in the body of the text. Finally, we have made the appropriate changes to the Tables section and added a section listing the additional files that were previously referred to as table 1 thru 6.

Regarding the surveillance ontology, which should be (in my opinion) the focus of this paper, I find what is shown not convincing. All the results are expressed on Figure 1 that displays which classes, relations, instances were chosen.

Actually, the true results are contained in the PHSkb database itself, not in the figure. Figure 1 is a gross over-simplification and a poor substitute for the richer knowledge representation contained in the PHSkb. It is impossible to express in Figure 1, without making it illegible, all of the relationships between concepts expressed in the knowledgebase. We refer the reviewer to the actual knowledgebase now available at the FTP site. Please see related changes to the manuscript described below under the next comment.

For instance in the class “Infectious agent “ and to my surprise, 6 instances are given (name, SNOMED code, Bioterror agent, associated disorder, jurisdiction reportable in). Rather I would expect subclasses (such as viral, bacterial, parasitic diseases), or at least instances allowing to enter such concepts.

Only major classes, their attributes, and broad relationships between classes are included in Figure 1. The class hierarchy is not expressed in the figure. Within the knowledgebase, taxonomic knowledge is expressed in the class hierarchy, rather than the attributes specified in Figure 1. Therefore, just as in SNOMED, one can infer that a particular organism is bacterial, viral, etc, from the hierarchical classification. Whether or not to represent knowledge as part of the class hierarchy or as an attribute of a class member is a design decision with no ‘correct’ answer. It is a matter of design preference. Ultimately our design decisions must be tested for their inference capability, however, we
chose to express taxonomic knowledge in a manner consistent with SNOMED, within the class hierarchy. This is described in the Construction and content section, Class hierarchy subsection.

To further clarify the purpose and content of the figure, the first line of the Construction and content section, Domain information model subsection has been modified from:

> An information model was created for the notifiable disease domain (figure) that depicts the major root classes in the ontology and their relation to one another.

To:

> An information model was created for the notifiable disease domain (Figure 1) that depicts the major root classes in the ontology, attributes for each class, and relationships between classes. By necessity, the figure is an oversimplification and does not account for the hierarchical classification within classes or identify all the links between classes. The reader is referred to the PHSkb for the full knowledge representation.

**By contrast, two instances deal with bioterrorism, which is an important problem but not to the point to be 2/6 of the instances of the class ‘infectious agent’**.

With due respect, we believe the reviewer is confusing instances with attributes. The 6 items in Figure 1 under infectious agents are attributes not instances of the class. As can be seen from former Table 3 (now additional file 3) and as described in the Class hierarchy subsection of the text, there are more than 174 instances in the infectious agent class, arranged hierarchically and not indicated in the figure. In contrast, knowledge related to bioterrorism agents does not have a clear hierarchical classification and is not contained in SNOMED. Therefore, we chose to represent this knowledge as an attribute or ‘slot’ in Protégé. The fact that bioterrorism related attributes account for 2/6 of the attributes, has no bearing on the overall adequacy of the knowledge representation, since many aspects of the knowledgebase cannot be expressed in figure 1. The reviewer is referred to the actual knowledgebase for greater details regarding the knowledge representation. To further clarify the distinction between attributes and instances, we have modified the first sentence of the second paragraph of the subsection on Content and relation to vocabulary standards to read, **“Additional attribute data for each class instance were then entered into the database (Figure 1).”**

Similarly, I would put ‘jurisdiction reportable in (?)’ in the class ‘Jurisdiction’ and make a link from Infectious agent to the class Jurisdiction. The same remarks apply to the other classes.

For purposes of legibility, not all links contained in the knowledgebase are indicated in the figure. We have attempted to clarify this in the text as previously stated. Such links between infectious agents and the jurisdictions where they are reportable, currently exist in the PHSkb. Currently, one can query the knowledgebase to determine what infectious
agents are reportable within a given jurisdiction, and conversely, what jurisdictions require the reporting of a given infectious agent. The same is true of reportable diseases, substances, findings, and procedures. We refer the reviewer to the knowledgebase itself and urge caution in interpreting the meaning of a particular attribute from Figure 1 based only on its name.

Note also that the entire paper deals only with the surveillance of infectious disease and not for instance with that of environmental diseases. This should be stated in the title, and abstract.

The knowledgebase scope is described in the first sentence of the Construction and content section and “includes diseases, conditions, or other events that are reportable in one or more reporting jurisdictions in the United States.” Thus the domain is neither purely infectious diseases, nor exclusive of environmental diseases. If a disease is reportable in at least one jurisdiction, it’s in-scope. If it’s not reportable in at least one jurisdiction, it’s out-of-scope (except to construct the class hierarchy as described in the manuscript). Indeed, table 2 includes several non-infectious or environmental diseases such as lead poisoning, pesticide poisoning, malignant neoplastic disease, etc. The other tables also include numerous non-infectious concepts. Therefore, it is our belief that the current title and abstract accurately reflect the scope of the knowledgebase content and appropriately highlight notifiability as the distinguishing characteristic which defines the domain. No changes were made to the manuscript in response to this comment.

Note finally that PHSkb was not downloadable at the FTP site indicated (when I tried).

The PHSkb files are now available for download at the FTP site http://ftp.cdc.gov/pub/epodphsi/PHSkb and the Availability and requirements section of the manuscript has been updated to reflect the current file location.

- Major Compulsory Revisions: Focus on the ontology. Give an in depth description of the chosen knowledge representation, and discuss the choices made.

The PHSkb ontology and knowledgebase has always been our primary focus and publishing it is our primary objective. The manuscript was formatted and submitted as a database article with the primary objective of describing the methods used to create the knowledgebase. The manuscript was structured in accordance with the instructions to authors for database articles to include sections for background, construction and content, utility, and discussion. Within these sections, authors are further instructed to include information on, “the perceived need for the database…stating its potential value…information on data sources…intended uses of the database…planned future development…importance and relevance of the database.” In summary, we believe the manuscript content is consistent with the instructions we were provided for submitting
The manuscript is meant to complement, not substitute for the ontology/knowledgebase, which is our primary focus. No additional changes were made to the manuscript.

- Minor Essential Revisions: Delete Tables 1 to 6 which must be placed in a public health bulletin, not in a public health informatics paper.

As previously mentioned, Table 1 provides some of the raw data used to instantiate the knowledgebase, and tables 2 thru 6 are artifacts produced from the knowledgebase. The advantage of the tables is that they provide a quick view of the breadth of the domain, without requiring readers to download special software to access the knowledgebase files. Under circumstances of a hardcopy printed journal, we would not consider submitting such lengthy tables for publication. However, with an all electronic journal, the tables can be provided to the reader as uploaded additional files. Tables 1 thru 6 are not meant to appear in the body of the text, but rather to be available as additional supporting files. We apologize for any confusion regarding this distinction. We have made changes throughout the text to refer to these as additional file 1 thru 6 instead of table 1 thru 6. In addition, we have renumbered table 7 to table 1, and included it in the body of the text. Finally, we have made the appropriate changes to the Tables section and added a section listing the additional files that were previously referred to as table 1 thru 6. We leave it to the editor’s discretion regarding whether or not the former tables 1-6 should be included in a final published manuscript as additional files. An alternative would be to make these additional files available in a folder for download from the FTP site used to access the PHSkb files.

Finally, we have made minor cosmetic changes to the title page, competing interests, author contributions, and references section to comply with requirements specified in the manuscript formatting checklist.

Thank you for your consideration of this revised manuscript.

Sincerely your,

Timothy Doyle, MPH