Title: A novel data-driven workflow combining literature and electronic health records to estimate comorbidities burden for a specific disease: a case study on autoimmune comorbidities in patients with celiac disease

Authors:

Jean-Baptiste Escudié (jean-baptiste.escudie@aphp.fr)
Bastien Rance (bastien.rance@aphp.fr)
Georgia Malamut (georgia.malamut@aphp.fr)
Sherine Khater (sherine.khater@aphp.fr)
Anita Burgun (anita.burgun@aphp.fr)
Christophe Cellier (christophe.cellier@aphp.fr)
Anne-Sophie Jannot (annesophie.jannot@aphp.fr)

Version: 2 Date: 18 May 2017

Author’s response to reviews:

MIDM-D-16-00292R1

A data-driven workflow using literature and electronic health records to phenotype comorbidities for a specific disease: application to autoimmune comorbidities in patients with celiac disease.

Jean-Baptiste Escudié, M.D.; Bastien Rance, Ph.D.; Georgia Malamut, M.D., Ph.D.; Sherine Khater, M.D.; Anita Burgun, M.D., Ph.D.; Christophe Cellier, M.D., Ph.D.; Anne-Sophie Jannot, M.D., Ph.D.

BMC Medical Informatics and Decision Making

Dear Editor,

We thank you for giving us the opportunity to submit a revised version of our manuscript now entitled A novel data-driven workflow combining both literature and electronic health records to estimate comorbidities burden for a specific disease: a case study on autoimmune comorbidities
in patients with celiac disease." (MIDM-D-16-00292). We have changed the title to more clearly explain that this is a case study application of a process for supporting the mining of EHRs by restricting the set of conditions to look for in the records, as requested by reviewer 1.

In this new version of the manuscript, we have made most of the changes requested by the reviewers (see below for a point-by-point response). All changes to the manuscript are indicated using track changes.

We wish to thank the reviewers for their helpful and constructive comments.

We hope that this revised version fulfils all requirements for a publication in BMC Medical Informatics and Decision Making.

With best regards,
Dr. Jean-Baptiste Escudie and Dr Anne-Sophie Jannot

Point by point reviewer response:

Adam Dunn (Reviewer 1):

Thank you for the opportunity to review this manuscript. The authors describe the application of what is broadly a literature discovery and EHR co-morbidity mining pipeline to autoimmune conditions associated with celiac disease. The research is interesting but I felt that the paper should be restructured and rewritten to more accurately reflect what was done. Currently the major flaw is that it describes the contribution as the data-driven workflow but the methods and results relate to what was discovered in a specific application, rather than an evaluation of the
performance and value of the data-driven workflow. Several of my comments below relate directly to this problem. The quality of the writing was high.

Response to reviewer: Thank you for all your comments.

1. Perhaps modifying the title to more clearly explain that this is a case study application of a process for supporting the mining of EHRs by restricting the set of conditions to look for in the records. I also immediately thought that the most interesting aspect of this research would be in looking at the "attention" in the literature versus the "incidence" discovered from the EHRs. To do this properly, however, I would expect to see a much broader range of comorbidities examined.

Response to reviewer: We modified the title accordingly: “A novel data-driven workflow combining both literature and electronic health records to estimate comorbidities burden for a specific disease: a case study on autoimmune comorbidities in patients with celiac disease.” We clarify along the text that we only focused on autoimmune diseases. For instance, the overview starts now as follows: “We first selected the list of autoimmune diseases in MeSH® terminology. We then restricted this list to the most frequent autoimmune diseases associated with CD from in the literature, based on the number of co-occurrences of MeSH® terms in MEDLINE®.” Regarding your point on the fact that the most interesting aspect of this research would be in looking at the "attention" in the literature versus the "incidence" discovered from the EHRs, we have now a dedicated paragraph on the correlation between prevalence estimates in our cohort and co-occurrence ranking in the literature. We also add in the discussion section the following sentence: “The novel approach mining literature presented in this study enabled to identify relevant comorbidities as attested by the fact that the attention from the literature was coherent with the prevalence found both in the literature and in our cohort”

2. Abstract: The background is long and doesn't match the methods and results, which describe only *how* the pipeline was applied to generate results for a specific condition and the results. There is no description of whether the approach produces a valid result or a statistical test confirming that the comorbidities identified through this process are either (a) the same as would be produced using an alternative method but automated (answering "why use the literature"); or (b) more accurate than other automatic approaches to identifying comorbidities from EHRs.
Response to reviewer: We shorten the abstract background and the latter is now more focused on the objective of our novel approach. We added information on how the performance of our approach is assessed, i.e. we added in the abstract: “The performance of our approach was assessed by estimating the correlation between prevalence estimates in our cohort and co-occurrence ranking in the literature for the 15 selected auto-immune CD comorbidities.”

3. It wasn't until about half way through the manuscript that I realised that the method was used to find a very restricted set of co-morbidities. I thought it would use a MeSH subset of *all* conditions and attempt to find those. From what I understand, it only selects from auto-immune conditions and does not explain or test how the method generalises.

Response to reviewer: We clarified the corresponding method subsection which is now entitled “Selection of a restricted set of auto-immune diseases using MeSH® co-occurrence file”. We also clarified its content by adding the following sentences: “We arbitrarily restricted the list for this study to the 15 autoimmune diseases the most co-occurring with CD. This process allows to restrict to a domain of comorbidities using MeSH hierarchy and to a subset of comorbidities using number of co-occurrences.” We added in the discussion section some explanation on how this method generalizes: “Our method is flexible as domain restriction using MeSH® hierarchy and limiting the number of results with the number of co-occurrence are both optional, although we haven't evaluated this method without these two types of restrictions.”

4. Background: I think the title on Page 4 is the wrong place to have the title.

Response to reviewer: We deleted the title on page 4.

5. Background: Reference [1] does not appear to support the statement that "Today in all major hospitals Clinical Data Warehouses gather information"

Response to reviewer: we modified this sentence and the reference accordingly: “Today in all major hospitals Clinical Data Warehouses (CDW) gather information on hundred thousands of patients, collected from Electronic Health Records (EHRs). These CDWs can be used to phenotype comorbidities as in our institution [1]”
6. Background: It might be worth trying for a more directed background, where the first paragraph introduces the problem (identifying comorbidities from EHRs), the second paragraph evaluates the existing approaches to doing this, an optional third paragraph that reviews literature discovery and any examples that might have been used in concert with EHRs, a paragraph on comorbidities in celiac disease, the next paragraph explains the rationale behind the approach that has been chosen, and the final paragraph states the main aims and objectives of the paper (e.g. a case study for one condition and one class of comorbidities).

Response to reviewer: Thank you for this proposal. We modified the background according to this proposal and the background now follows the proposed scheme:

- the first paragraph introduces the problem (identifying comorbidities from EHRs)

- the second paragraph evaluates the existing approaches: EHR sources that can be mined and phenotyping algorithm

- the third paragraph deals with comorbidities in celiac disease

- the fourth paragraph explains the rationale behind the approach that has been chosen “To the best of our knowledge, there is no clear review on the most prevalent set of autoimmune comorbidities associated to CD, while there is a need to phenotype autoimmune comorbidities burden in CD patients to enable further stratification of patients’ profiles. (…)”
- the final paragraph states the main aims and objectives of the paper, i.e. “identifying relevant autoimmune comorbidities in celiac disease and to phenotype for these autoimmune comorbidities the population of adult CD patients”.

7. Background: The paragraph explaining what is known about comorbidities in celiac disease is useful and an important paragraph to include. It might be worth arguing that besides the apparent variability measured in different places, to the best of your knowledge there has not been a clear synthesis of these studies published. It would be fine to say that this study *adds* to this literature by identifying comorbidities in a set of 741 new patients, and using a novel approach to the identification. To reiterate, there is nothing wrong with this as an aim.

Response to reviewer: We add the following sentence in the background section: “To the best of our knowledge, there is no clear review on the most prevalent set of autoimmune comorbidities associated to CD, while there is a need to phenotype autoimmune comorbidities burden in CD patients to enable further stratification of patients’ profiles.” We also added as one of the study objective: ”to assess its [workflow] performance in this context by assessing quantitatively whether literature-based knowledge was correlated to EHR-based extracted knowledge regarding autoimmune CD comorbidities” This performance assessment is now in a dedicated paragraph in the result section.

8. Methods: In the study population section, aim to be more precise so it is clear what the study data encapsulate. The total number of patients and documents could be written here.

Response to reviewer: We change the population section to make clearer what the study data encapsulate: we now first describe patients’ inclusion from the list of patients with celiac disease maintained by clinicians and then we describe how we completed this list of patients. We also add the following sentence (that was previously in the results section): “The study population counted 741 patients and a corpus of 6,340 clinical reports.”
9. Methods: In the manuscript, it is very hard to understand how the co-occurrence in the literature is used to support the phenotyping, or what the actual purpose of using the literature is. I thought perhaps the aim of the study was to examine how these are different, to support hypothesis generation, and to design more pragmatic trials. After reading the manuscript, it now seems as though there may have been no reason to use the literature at all, and all that was needed was to select 15 auto-immune diseases and leverage existing terminologies to support phenotyping.

Response to reviewer: in our study, literature serves both purposes:

- to design more pragmatic studies, not relying on one or two experts’ opinion;

- to select a restricted set of disease when there is no clear review on the most prevalent set of comorbidities.

We clarified the subsection “Literature-based selection of autoimmune diseases” of the discussion accordingly and added the following sentences: “To the best of your knowledge there was no clear synthesis of major CD autoimmune comorbidities. (…) This method allows to design more pragmatic studies, not relying on one or two experts’ opinion.”

10: Pointing the reader to FASTVISU and skipping over the process for identifying comorbidities confused me in the first instance - I didn't realise that this was a manual process until I read through it a second time. Perhaps an improved flow diagram might help. Then I wondered - why do you need the literature discovery at all? Why not just highlight anything that matches any of the synonyms from the terminologies in the software and let people confirm it manually?

Response to reviewer: Actually, FASTVISU highlights only synonyms for the selected set of auto-immune comorbidities. FASTVISU does not encapsulate terminologies and is designed to vote for a given set of predefined conditions. We clarified the text accordingly: “FASTVISU highlights terms with an entity recognition module based on regular expressions (e.g., the pattern
(\bdiab\w+\b to match for diabetes) defined by the user and approximate syntax matching techniques. For the 15 selected auto-immune comorbidities, a set of regular expression was defined and two trained physicians reviewed the entire corpus of clinical narratives using FASTVISU based on this set of regular expression to validate the presence/absence of each of the selected auto-immune diseases for each patient.”

11: Results: I don't think you need to capitalise words in the table title - sentence capitalisation is fine.
Response to reviewer: we changed the tables’ titles accordingly.

12. Results: When reporting Cohen's kappa, also report the raw agreement (at least). This is a key performance indicator in the process, because the FASTVISU software is enabling faster and more consistent labelling of comorbidities by humans. How long would it have taken them without FASTVISU and would the agreement have been lower? What if FASTVISU was set up with different (fewer) terms that were highlighted?

Response to reviewer: We now also reported raw agreement: we added in the results section: “Readers voted on 466 items. More specifically, 465 patients out of the 741 included patients had no highlighted terms and 466 autoimmune disease items had at least a highlighted term on the 276 remaining patients. For 140 items, voters both approved that the patient suffered from this disease; for 304 items, readers both disapproved that the patient suffered from this disease and for 22 items, readers mutually disagreed. Therefore, inter-reviewer agreement for auto-immune disorder identification in narrative reports was excellent, with a Cohen’s kappa value of 0.89.”. Unfortunately, we did not compare the performance of fastvisu with another method as there is no gold-standard method for such extraction.

13. Results: Figure 3 is interesting, I believe it shows the location in the EHR where the comorbidity was identified. The implications of these results are also described in the discussion, and I thought this was a useful and important contribution.
Response to reviewer: Thank you for your comment.

14. Discussion: I think the literature discovery results are first described in the Discussion section and appear to be out of place.

Response to reviewer: We reorganized discussion section to be in line with results section. The first paragraph of the discussion deals now with the external validation of prevalence estimates

15. Discussion: I think the comparison with other prevalence estimates should be included in the results (just the comparison not the explanation). I think this comparison is an important evaluation of the performance of the method in the context of the specific dataset used, and the discussion section should only explain/speculate about why the results are different:

Response to reviewer: The result section now includes prevalence estimates from the literature (table 3) and the discussion section only speculate about why the results are different.

Overall, I think the manuscript would benefit from a clearer structure, and clearer explanations of (a) the places in the EHR where comorbidities are extracted; (b) comparing the use of different terminologies and synonyms to undertake the phenotyping; (c) comparison of the prevalence in the population versus the "attention" in the literature. At the moment it only really describes the steps without explicitly measuring the performance of the method or comparing the approach to alternatives.

Response to reviewer: The manuscript now encloses some explanations on the places in the EHR where comorbidities were extracted and the use of regular expression (synonyms) to undertake the phenotyping: “For the 15 selected auto-immune comorbidities, a set of regular expression was defined and two trained physicians reviewed the entire corpus of clinical narratives using FASTVISU based on this set of regular expression to validate the presence/absence of each of the selected auto-immune diseases for each patient.” We now are more focused on the performance for the comparison of the prevalence in the population versus the "attention" in the
literature. We added in the method section the following sentence: “Finally, to assess the performance of literature-based comorbidities selection, we estimated the correlation between MeSH® and diseases prevalence, using Spearman’s correlation coefficient between number of publications indexed with MeSH® terms for both CD and an autoimmune disease, and the prevalence the corresponded disease obtained from our EHR extraction.”

Manoj Mammen (Reviewer 2): Overall a well-written article, that is pertinent and accessible to a wide audience. The generic workflow the authors has described, appears to be novel and able to phenotype subjects for related co-morbidities of a specified disease.

Response to reviewer: Thank you for your comment.