Author’s response to reviews

Title: Rare Disease Knowledge Enrichment through a Data-Driven Approach

Authors:

Feichen Shen (Shen.Feichen@mayo.edu)
Yiqing Zhao (zhao.yiqing@mayo.edu)
Liwei Wang (Liwei.Wang@mayo.edu)
Majid Mojarad (Majid.Mojarad@mayo.edu)
Yanshan Wang (Wang.Yanshan@mayo.edu)
Sijia Liu (Liu.Sijia@mayo.edu)
Hongfang Liu (Liu.Hongfang@mayo.edu)

Version: 2 Date: 05 Jan 2019

Author’s response to reviews:

Response to Reviewers:

We thank the editor and reviewers for their valuable suggestions. In this revised version, we carefully revised the manuscript based on all the suggestions.

Below please find our point-to-point answers to comments. Figures, Tables, and References refer to the revised manuscript. The number of page indicated in this letter is counted starting from the main body of the manuscript. All the updated part in manuscript for this round was highlighted using red color.

Editor’s Comments:

1. The first step of the workflow is to mine the association rules. Selecting an appropriate value of support and confidence is essential, as these parameters drive the set of association rules that goes onto the next step of the workflow. It would be useful to better characterize how the
characteristics of the association rules depend on support and confidence. There are many options available, which include but is not limited to the following. A heatmap can show how a characteristic depends on support (x-axis) and (y-axis). The most common characteristic would be the number of rules. Another one would be the average size of the antecedent, and so on. I would invite the authors to show how support and confidence shape the characteristics of the association rules via such heatmaps or appropriate alternative forms of visualizations across the combinations of parameter values.

Response 0.1: We appreciate the editor’s comment on this. We have made a heatmap to characterize the associations. Specifically, as shown in Figure 3, x-axis indicates support value ranges from 1.24E-06 to 1.37E-05 and y-axis indicates confidence value ranges from 0.0005 to 1. The number of rules is represented by different colors. Our observation is that the number of rules gets decreased with the increment of both support and confidence.

Text change (Page 9, lines 191-194): We also characterized the association rules using a heatmap as shown in Figure 3. Specifically, x-axis indicates support value ranges from 1.24E-06 to 1.37E-05 and y-axis indicates confidence value ranges from 0.0005 to 1. From Figure 3, we observed that the number of rules get decreased with the increment of both support and confidence.

2. as shown by the authors, association rules can be visualized as networks. It would be useful to point out that network analysis can further characterize the association rules (c.f., https://link.springer.com/article/10.1007/s11227-016-1714-y) and reveal relations between diseases that may not have been easily apparent otherwise.

Response 0.2: We appreciate the editor’s comment on this. We have added this point into discussion section, claiming that some network analysis approaches could be incorporated with association rule mining to reveal hidden relationships among diseases in the future work.

Text change (Page 15, lines 282-284): Moreover, some network analysis approaches [57] will be incorporated to further characterize the association rules and reveal hidden relations among diseases.
Reviewer 2:

1. The first one is the inclusion of the Jaccard scores in Table 5 so that the reader can get an idea of how far each proposed disease is from the reference disease.

Response 2.1: We thank the reviewer for this point. We have added Jaccard scores for our approach (HPO-Orphanet+) and the eRAM. Since Phenomizer leveraged the information content (IC)-based score to rank diseases, we also introduced this ranking strategy in material section and added IC-based scores for the Phenomizer in Table 5.

Text change (Page 5, lines 97-100): Specifically, the HPO-Orphanet+ and eRAM ranked diagnostic candidates by the descending order of Jaccard similarity score [29], and the Phenomizer ranked diagnostic candidates by descending order of Information Content (IC)-based similarity score proposed in [18].

Table 5 with caption were updated (Pages 13-14)

2. The second one is the validation by either experts or the literature of the identified common diseases in the differential diagnoses of this table.

Response 2.2: We’d like to thank the reviewer for this point. In this revision, we manually checked scientific literature and online knowledgebase/materials to validate the associations mined by the HPO-Orphanet+. We found that 10 out of 15 diagnostic candidates were proved to be strongly associated with Hodgkin lymphoma. For example, lung adenocarcinoma has similar characterizations with Hodgkin lymphoma, and glomerulonephritis is a well-recognized complication of Hodgkin disease. Few evidences were detected for dilated cardiomyopathy, abdominal aortic aneurysm, degenerative polyarthritis, atrial fibrillation, and glaucoma from online materials and scientific literature, indicating that the associations mined from patients’ data provided new evidences for differential diagnosis of Hodgkin lymphoma.

In addition, for those novel disease-phenotype associations mined from data and cannot be validated by biomedical literature, online database or knowledge base, we will recruit domain
experts to provide a manual evaluation and curate the enriched knowledge base in the future work. We have added this point in discussion section.

Text change (Page 13, lines 246-253): In addition, based on literature and online material review, we found that 10 out of 15 diagnostic candidates were proved to be strongly associated with Hodgkin lymphoma based on similar comorbidities or complications [5, 40-48]. For example, lung adenocarcinoma has similar characterizations with Hodgkin lymphoma [46], and glomerulonephritis is a well-recognized complication of Hodgkin disease [49]. Few evidences were detected for dilated cardiomyopathy, abdominal aortic aneurysm, degenerative polyarthritis, atrial fibrillation, and glaucoma from online materials and scientific literature, indicating that the associations mined from patients’ data provided new evidences for differential diagnosis of Hodgkin lymphoma.

Text change (Pages 14-15, lines 271-275): Moreover, for those novel disease-phenotype associations mined from data and cannot be validated by biomedical literature, online database or knowledge base, we will recruit domain experts to provide a manual evaluation and curate the enriched knowledge base in the future work. More evaluation metrics (e.g., precision, recall, and F-measure) will be applied based on experts’ judgements.

3. Another issue comes from the sentence in lines 206-208. How is it possible that the degree of certain diseases gets lower after enriching the graph? Are some phenotypes removed when enriching the graph?

Response 2.3: We appreciate the reviewer’s comment on this. This is a misinterpretation we made in the last version. First of all, all those certain diseases got an increased degree. However, some other diseases received a larger increment of degree and resulting higher rankings. The message we wanted to deliver here is actually about the ranking of the degree. We have corrected this.

Text change (Page 11, lines 219-220): But some dominant diseases in HPO-Orphanet graph were not ranked high in the HPO-Orphanet+ graph.

4. Other minor comments are as follows:
4.1 Reference [17], what is this? PhD, Technical Report?

Response 2.4.1: We have corrected this.


4.2 The term performance is uncountable in the context it is used

Response 2.4.2: We have fixed this throughout the manuscript.

4.3 Line 204: combining two datasets -> combining the two datasets

Response 2.4.3: We have fixed this.

Text change (Page 11, line 217): combining the two datasets.

4.4 Line 211: diseases with high degrees -> diseases with the highest degrees

Response 2.4.4: We have fixed this.

Text change (Page 11, line 224): diseases with the highest degree.

4.5 Line 213: to compare performances -> to compare the performance
Response 2.4.5: We have fixed this.

Text change (Page 12, line 226): to compare the performance.

4.6 Line 247: it is not necessary to include references for bigrams and trigrams

Response 2.4.6: We have removed the references.

4.7 Line 271: naive Bayes -> Naïve Bayes

Response 2.4.7: We have fixed this.

Text change (Page 15, line 297): Naïve Bayes

4.8 Reference [47], what is this? Journal, conference?

Response 2.4.8: This reference is cited as reference [53] in this revised version. This is a conference paper and we have fixed this.

4.9 Many references are badly formatted, for example in many journals they do not include volume, number etc.

Response 2.4.9: We have reformatted all references. We will follow the guidelines and work with associate editors of the Journal to further refine our reference format.
Reviewer 4:

1. page 4 line 78. With the revision of the text to indicate that all clinical notes were used in the study, it is now clearer than the previous version on what data was used in the study. However, it would be much clearer if the authors can include the example types of EMR notes mentioned in the responses in the manuscript. That is, my suggestion is to add into the manuscript the text from the responses "including Consultant Notes (CON), … for each patient".

Response 4.1: We’d like to thank the reviewer for this point. We have added some example types of EMR notes in the manuscript.

Text change (Page 4, lines 78-81): All clinical notes during the years of 2010 to 2015 from Mayo Clinic EMR were used for the study, including Consultant Notes (CON), Subsequent Visit Notes (SV), Emergency Medicine Notes (EMV), Hospital Admission Notes (ADM), and so on. For each note type, we focused on the diagnosis section of the notes which summarizes problems for each patient.

2. page 8. line 164. So the ground truth was collected by merging the two files mentioned in the manuscript. I am not an expert on medical knowledge, should there be some form of manual check to make sure that the ground truth is truly correct?

Response 4.2: We appreciate the reviewer for this comment. The eRAM is a human annotated resource that records associations between diseases and phenotypes. According to the eRAM paper [16], all associations maintained by the eRAM were manually checked by domain experts, including the two files: “eRAM Integrated Phenotype.txt” and “eRAM Integrated Symptom.txt”. Therefore, in this study, we considered the associations provided by the merged files as correct ones to prepare the gold standard. We have highlighted this part in the revised manuscript.

Text change (Page 8, lines 170-171): Since those two files were manually annotated by domain experts, in this study, we considered the associations provided by the merged files as correct ones to prepare the gold standard.
3. page 8 lines 172, 173. To me, it is more common to calculate precision, recall and F1 measure as a way to evaluate a classification system. precision = TP/(TP+FP), recall = TP/(TP+FN), F1 = 2xPrecisionxRecall/(Precision+Recall). Furthermore, with F1 measure, it is possible to calculate statistical significant differences among different models. Consequently, the question on Figure 3 is whether there was significant differences among the three graphs.

Response 4.3: We appreciate the reviewer for this comment. First of all, this study is not a classification problem. Our purpose is to acquire new knowledge using a data-driven approach. The information retrieved from data might be or might not be recorded in an existing database/knowledge base. Therefore, in this study, we first used increment of explanatory power (IEP) to quantify how much new information been discovered. In addition, we took the eRAM as a gold standard and used specificity and sensitivity as two metrics to conduct the evaluation based on how much information can be found from the gold standard. Although the overlap between the discovered knowledge and the gold standard is limited, from Figure 4, we can still observe that the enriched knowledge yielded significant performance improvement in terms of sensitivity, indicating that it is a benefit of combining two resources for a better differential diagnosis.

Combining this comment with comment #6, in the future, we will recruit some domain experts to provide a further validation on the newly discovered knowledge (for those cannot be found in existing literature, database or knowledge base). More evaluation metrics will be also applied by then. We have added some discussions in the manuscript.

Text change(Pages 14-15, lines 271-275): Moreover, for those novel disease-phenotype associations mined from data and cannot be validated by biomedical literature, online database or knowledge base, we will recruit domain experts to provide a manual evaluation and curate the enriched knowledge base in the future work. More evaluation metrics (e.g., precision, recall, and F-measure) will be applied based on experts’ judgements.

4. page 10, line 198. it states that "the density for the HPO-Orphanet+ graph was in between HPO-Orphanet graph and EMR graph", but table 3 shows that the density for the HPO-Orphanet+ graph, HPO-Orphanet graph and EMR graph is 0.005, 0.006 and 0.007, respectively. So HPO-Orphanet+ graph has the lowest density value. Did I miss anything?
Response 4.4: We appreciate the reviewer’s comment. The results are correct but we misinterpreted the results in last version. HPO-Orphanet+ do had the lowest density based on our observation. According to Equations (7) and (8), for a given graph G, Density(G) is calculated by $\Delta \bar{G}(|v|-1)$. Although the HPO-Orphanet+ held the highest average degree, since vertices got enriched, resulting a relative lower density for the HPO-Orphanet+.

Text change (Pages 10-11, lines 209-213): The density for the HPO-Orphanet+ graph was the lowest among all graphs. According to Equations (7) and (8), for a given graph G, Density(G) is calculated by $\Delta \bar{G}(|v|-1)$. Although the HPO-Orphanet+ held the highest average degree, since vertices got enriched, resulting a relative lower density for the HPO-Orphanet+.

5. page 11 table 4 and page 12 table 5. Both tables list 15 different candidates for each graph respectively. I am confused on the meaning of these candidates. Are the ones generated from the graph as the ones that are most similar to a given rare disease such as Hodgkin lymphoma? Since there is ground truth on the correct association, should they be checked on whether the candidates are correct or not?

Response 4.5: We appreciate the reviewer’s comment. Table 4 shows diseases with the highest degree in three different graphs. Table 4 only describes graph characterizations (e.g., diseases with the highest average degree) and NOT about differential diagnostic candidates. While Table 5 displays top 15 differential diagnostic candidates for Hodgkin lymphoma. In this revision, we manually checked scientific literature and online knowledgebase/materials to validate the correctness of these candidates. We found that 10 out of 15 diagnostic candidates were proved to be strongly associated with Hodgkin lymphoma. For example, lung adenocarcinoma has similar characterizations with Hodgkin lymphoma, and glomerulonephritis is a well-recognized complication of Hodgkin disease. Few evidences were detected for dilated cardiomyopathy, abdominal aortic aneurysm, degenerative polyarthritis, atrial fibrillation, and glaucoma from online materials and scientific literature, indicating that the associations mined from patients’ data provided new evidences for differential diagnosis of Hodgkin lymphoma. We also added scores in Table 5 for different tools to give audiences some intuitions about the ranking.

Text change (Page 13, lines 246-253): In addition, based on literature and online material review, we found that 10 out of 15 diagnostic candidates were proved to be strongly associated with Hodgkin lymphoma based on similar comorbidities or complications [5, 40-48]. For example, lung adenocarcinoma has similar characterizations with Hodgkin lymphoma [46], and
glomerulonephritis is a well-recognized complication of Hodgkin disease [49]. Few evidences were detected for dilated cardiomyopathy, abdominal aortic aneurysm, degenerative polyarthritis, atrial fibrillation, and glaucoma from online materials and scientific literature, indicating that the associations mined from patients’ data provided new evidences for differential diagnosis of Hodgkin lymphoma.

Table 5 and caption were updated (Pages 13-14)

6. so figure 3 shows some performance score of the system, and figure 4 shows an interface of the system. Have the authors conducted any evaluation, even just initial ones, with domain experts to see if the results generated by the system make sense or useful?

Response 4.6: We appreciate the reviewer for this comment. The sensitivity and specificity for generating differential diagnostic suggestions for Hodgkin lymphoma was made as an evaluation to show the performance among three different bipartite graphs (as shown in Figure 4 in this revision). In this evaluation, we took the candidates generated by the eRAM as the gold standard. This gold standard was manually generated by domain experts according to the eRAM paper. From Figure 4, the increment of sensitivity actually indicates the performance improvement of combined EMR and knowledgebase, and that shows the benefit of combining two resources for differential diagnosis suggestions. In addition, as shown in Response 4.5, in this revision, we have manually checked scientific literature and online knowledgebase/materials to validate the correctness of the candidates generated by the HPO-Orphanet+. We found that 10 out of 15 diagnostic candidates were proved to be strongly associated with Hodgkin lymphoma.

In the future, we will recruit some domain experts to provide a further validation on the newly discovered knowledge (for those cannot be found in existing literature, database or knowledge base). We have added some discussions in the manuscript.

Text change(Pages 14-15, lines 271-275): Moreover, for those novel disease-phenotype associations mined from data and cannot be validated by biomedical literature, online database or knowledge base, we will recruit domain experts to provide a manual evaluation and curate the enriched knowledge base in the future work. More evaluation metrics (e.g., precision, recall, and F-measure) will be applied based on experts’ judgements.