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: 1 Date: 16 Nov 2018

Author’s response to reviews:

Response to Reviewers:

We thank all reviewers for their valuable suggestions. In this revised version, we have made significant updates of this manuscript. We carefully revised the manuscript based on reviewers’ suggestions.

Below please find our point-to-point answers to comments from the reviewers. 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 was highlighted using red color.

Reviewer 1:

1. The study utilized clinical notes from EMR. What note types were included? E.g. HP, nursing flowsheet, discharge … or all EMR note types?
Response 1.1: We thank the reviewer’s valuable comments. In this study, we used all types of EMR notes, 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.

Text change (page 4, line 78): All clinical notes during the years of 2010 to 2015 from Mayo Clinic EMR were used for the study.

2. What constitute a phenotype in this study? All terms in HPO? How many phenotypes?

Response 1.2: We appreciate the reviewer’s comments. In this study, we used a previously developed HPO annotation pipeline to detect phenotypic term from clinical narratives. Specifically, we considered all terms recorded by the HPO as phenotypic terms and looked up the HPO as a phenotype dictionary. As a result, we extracted 2,808 unique phenotypic terms from clinical notes that also can be found from HPO. We have added this in Material section.

Text change (pages 4-5, lines 83-88): Specifically, the HPO was used to identify rare diseases and their phenotypic characterization mentioned in clinical narratives, and the UMLS was utilized to detect synonyms for any phenotypic terms. We limited our annotation to sections containing problems and diagnoses where 38,097 patients were found to have at least one diagnosis of a rare disease. Leveraging this pipeline, we extracted 2,808 unique phenotypes from notes and 9,292,969 phenotype-disease associations in total, from which 164,792 associations were related to 1,449 rare diseases and the rest were generated from 13,821 common diseases.

3. Typically, support and confidence are used together in association rule mining to determine significance of the rules learned. It usually requires user-specified minimum support and minimum confidence at the same time. A minimum support threshold is to find all frequent itemset and a minimum confidence is to apply to these frequent itemset to form rules. Using them as separate measures for discovering interesting rules is strange.

Response 1.3: We thank the reviewer for the insightful comment. We have revised our approach to select top associations as suggested. Specifically, we first use support to choose the frequent itemsets. In order to select the optimal threshold for support, we chose the average support as the
minimum support (5E-06). We then used confidence to choose the optimal associations out of those selected frequent itemsets. Similarly, we chose the average confidence as the minimum confidence (0.05). Figure 2 shows the curves do not have much fluctuation with values lower than the threshold. As a result, 13,742 associations (see additional file 1) were identified.

Text change (page 9, lines 177-182): To select the optimal associations, we set average support score 5E-06 as the threshold to first select 31,211 frequent itemsets and we then set average confidence score 0.05 as the minimum confidence to finalize 13,742 associations (see additional file 1). To further validate the selection of thresholds, as shown in Figure 2, we found that both support and confidence value didn’t have much fluctuation after dropping below their average values (the threshold point is marked on the curve).

All the consequent results were also updated accordingly in Tables 2-5, Figure 3 and relevant descriptions.

4. Phenotype-disease associations in SemMedDB has rankings? How were the rankings in SemMedDB determined?

Response 1.4: We’d like to thank reviewer’s comments on this point. We have revised the manuscript and no longer used the DCG ranking as the evaluation metric. Therefore, ranking was not considered for SemMedDB anymore. We only used SemMedDB to quantify the explanatory power for the enrichment as shown in page 10.

5. Using disease co-occurrences in EMR as a gold-standard for validating correctness of suggested disease by Phenomizer and D3N may not be a good approach as: (1) disease co-occurrence in a note does not guarantee a positive association; (2) D3N is EMR-enriched so the evaluation between phenomizer and D3N against the disease co-occurrences in EMR is biased toward D3N.

Response 1.5: We thank the reviewer’s valuable comments. According to the reviewer’s feedback, we have redesigned our evaluation and made significant changes on the evaluation.
For the first question, we no longer use co-occurrences in EMR as a gold-standard. Instead, we adopted an existing rare disease encyclopedia, named “eRAM”, to generate gold standard for this evaluation. Since data generated by eRAM is coming from both literature and EMR, it is unbiased and fair to compare the eRAM with our tool. Detailed on how to generate such gold standard is showed below.

Text change (page 8, lines 161-168): To prepare the experiment, for any disease to be tested, we used the three aforementioned graphs to rank suggested diseases with descending order of Jaccard similarity score. We combined two disease-phenotype association files namely “eRAM Integrated Phenotype.txt” and “eRAM Integrated Symptom.txt” provided by the eRAM. Based on 5,356 curated diseases and their associated phenotypes/symptoms obtained from the files, we calculated Jaccard similarity score between each pair of diseases using Equation 4. Such disease-disease similarity was used as a gold standard on differential diagnosis. In this evaluation, we first validated diagnostic candidates generated by the three bipartite graphs using the gold standard. We then compared the top 15 differential diagnostic candidates generated by the HPO-Orphanet+ graph, Phenomizer, and eRAM.

For the second question, first of all, we no longer named the network as D3N since it might confuse audiences. In particular, the aim of this study is to validate the effectiveness of the enriched knowledge base versus the original one. Therefore, we compared performances on disease recommendations for three bipartite graphs leveraging the gold standard provided by the aforementioned eRAM: original HPO-Orphanet Graph, EMR Graph, and Enriched HPO-Orphanet Graph (the new name is HPO-Orphanet+). We observed that the HPO-Orphanet+ graph shows the highest sensitivity for detecting the right similar diseases according to the gold standard, while using the original HPO-Orphanet based graph yields the lowest sensitivity (As shown in Figure 3).

Text change (page 12, lines 218-226): Sensitivity and specificity for generating differential diagnostic suggestions for Hodgkin lymphoma with different graphs is shown in Figure 3. The HPO-Orphanet+ graph shows the highest sensitivity for detecting the right similar diseases according to the eRAM gold standard, while using the HPO-Orphanet graph yields the lowest sensitivity. In addition, specificity does not show significant differences among three graphs, indicating that all of them have similar performances on rejecting non-relevant diseases for Hodgkin lymphoma. In general, we observed that the HPO-Orphanet+ graph enriched the existing rare disease knowledge resources and thus be able to provide better diagnostic suggestions. A web-based tool was implemented to visualize diagnostic suggestions and Figure 4.
shows an example of this differential diagnostic decision aid interface by considering Hodgkin lymphoma as center node.

In addition, we compared the top 15 differential diagnostic suggestions among the HPO-Orphanet+, eRAM, and Phenomizer using Hodgkin lymphoma. Since many rare diseases are commonly misdiagnosed as common diseases, it is essential to link common and rare diseases at the early time of diagnosis to assist diagnosis decision making. Compared to the Phenomizer and eRAM, we found that the HPO-Orphanet+ graph is more capable of detecting such associations between rare and common diseases. The top 15 diagnostic suggestions for the HPO-Orphanet+, eRAM, and Phenomizer can be found in Table 5.

Text change (page 12, lines 227-237): Table 5 shows top 15 differential diagnostic candidates for Hodgkin lymphoma between the HPO-Orphanet+ graph and two existing diagnostic tools (Phenomizer and eRAM). The HPO-Orphanet+ graph identified 46.7% (7 out of 15) common diseases and 53.3% (8 out of 15) rare diseases. Specifically, chronic obstructive airway disease, diabetes mellitus, atrial fibrillation, glaucoma, coronary heart disease, degenerative polyarthritis, and chronic kidney insufficiency are common diseases that share the most similar phenotypes with Hodgkin lymphoma, which are considered to be potential candidates for misdiagnosis of Hodgkin lymphoma. While differential diagnostic candidates provided by the Phenomizer are all rare disease. Similarly, the eRAM generates 93.3% (14 out of 15) rare diseases but only 6.7% (1 out of 15) common diseases. Since many rare diseases are commonly misdiagnosed as common diseases, it is essential to link common and rare diseases at the early time of diagnosis to assist in diagnostic decision support. Compared to the Phenomizer and eRAM, the HPO-Orphanet+ graph is more capable of detecting such associations.

Reviewer 2:

1. One of the most confusing points of the paper is the introduction of D3N. A D3N is mainly aimed at giving a ranking a diseases given a set of observed symptoms/phenotypes. It must be said that authors provide no reference to existing D3N approaches. Broadly speaking D3N aims at measuring the impact of each factor during the diagnosis decision making via a variety of methods like Bayesian estimation and/or machine learning methods (Jiang et al. 2017). In the proposed method, authors define a rather simple binary similarity measure (Jaccard) for comparing diseases in terms of their phenotypes. This similarity can be useful for evaluating
potential misdiagnosis, but not as a D3N since it requires assigning/learning weights to the phenotypes and their attributes.

Response 2.1: We thank the reviewer’s valuable comments. We agreed with reviewer’s points and have made significant updates on this.

First of all, we no longer named the network as D3N since it might confuse audiences. In particular, the aim of this study is to validate the effectiveness of enriched knowledge base versus the original one, therefore, we compared performances on differential diagnostic suggestions for three bipartite graphs: original HPO-Orphanet Graph, EMR Graph, and Enriched HPO-Orphanet Graph (the new name is HPO-Orphanet+) (List of abbreviations on page 14 details each of the name).

Introduction for the HPO-Orphanet+ was added in introduction section. Figure 1 was also updated accordingly.

Text change (page 4, lines 70-73): Here, we used the HPO annotation file named “phenotype_annotation.tab” accessed in July 2017 for association information between HPO terms and rare diseases in Orphanet [20]. These associations, which we referred to as HPO-Orphanet associations, were treated as rare disease knowledge resource in this study. We propose to enrich the HPO-Orphanet through mining association information between clinical phenotypes and diseases using EMR. Such enriched information, named as HPO-Orphanet+, can be used to link similar rare/common diseases and provide differential diagnostic decision aid at the point of care for rare disease diagnosis.

In addition, the evaluation focuses on evaluating disease similarity and potential misdiagnosed based on phenotypic characterizations provided by three different data sources. We have also added the relevant paper (Jiang et al. 2017) and discuss it in the discussion section. This work was able to calculate the probability of getting specific diseases from multisymptom naive Bayes algorithm. The third layer of “property” or multisymptoms is an interesting concept that may be involved in our future work.

Text change (page 14, lines 263-271): We compared the HPO-Orphanet+ with both the Phenomizer and eRAM in this study on differential diagnostic suggestions. Results showed that the HPO-Orphanet+ is capable of providing a diagnostic graph mixed with both rare and
common diseases, which has potential usage in rare disease differential diagnosis, especially for those rare diseases sharing similar symptoms with common diseases. In the future, we will upgrade the HPO-Orphanet+ by mining disease-gene information from literature [11, 51]. In addition, one recent research proposed a novel idea by introducing the concept of “property” as a third layer in addition to traditional two-layer disease-phenotype relationship [52]. This study was able to calculate the probability of getting specific diseases from a multisymptom naive Bayes algorithm. The third layer of “property” or multisymptoms is an interesting concept that may be involved in our future work.

2. Another confusing point is the comparison to Phenomizer. Which is the purpose of this comparison? Since the evaluation is performed by using the EMR data as ground truth, it is obvious that the proposed method is going to work much better than Phenomizer. Indeed authors agree with this as in the conclusions they say that "the comparison actually indicate the graph of diagnosis support between ...". Apart from comparing to Phenomizer, authors must evaluate sensitivities/specificities for the three proposed configurations, namely: EMR data only, HPO-Orphanet data only and the combined graph (Table 3). This comparison is neutral to the underlying ranking method (D3N vs. Phenomizer), providing thus proper evidence of the benefits of enriching KRs with EMR data.

Response 2.2: We thank the reviewer’s valuable comments. We agreed with reviewer’s points and have made significant updates on this. As we mentioned in Response 2.1, we have compared performances on recommending similar diseases for three graphs: original HPO-Orphanet Graph, EMR Graph, and Enriched HPO-Orphanet Graph (the new name is HPO-Orphanet+).

We no longer use co-occurrences in EMR as a gold-standard. Instead, we adopted an existing rare disease encyclopedia, named “eRAM”, to generate gold standard for this evaluation. Since data generated by eRAM is coming from both literature and EMR, it is unbiased and fair to compare the eRAM with our tool. Detailed on how to generate such gold standard is showed below.

Text change (page 8, lines 161-168): To prepare the experiment, for any disease to be tested, we used the three aforementioned graphs to rank suggested diseases with descending order of Jaccard similarity score. We combined two disease-phenotype association files namely “eRAM Integrated Phenotype.txt” and “eRAM Integrated Symptom.txt” provided by the eRAM. Based on 5,356 curated diseases and their associated phenotypes/symptoms obtained from the files, we
calculated Jaccard similarity score between each pair of diseases using Equation 4. Such disease-disease similarity was used as a gold standard on differential diagnosis. In this evaluation, we first validated diagnostic candidates generated by the three bipartite graphs using the gold standard. We then compared the top 15 differential diagnostic candidates generated by the HPO-Orphanet+ graph, Phenomizer, and eRAM.

As the reviewer suggested, we compared sensitivity and specificity for three graphs respectively. We observed that the HPO-Orphanet+ graph shows the highest sensitivity for detecting the right similar diseases according to the gold standard, while using the original HPO-Orphanet based graph yields the lowest sensitivity (As shown in Figure 3).

Text change (page 12, lines 218-226): Sensitivity and specificity for generating differential diagnostic suggestions for Hodgkin lymphoma with different graphs is shown in Figure 3. The HPO-Orphanet+ graph shows the highest sensitivity for detecting the right similar diseases according to the eRAM gold standard, while using the HPO-Orphanet graph yields the lowest sensitivity. In addition, specificity does not show significant differences among three graphs, indicating that all of them have similar performances on rejecting non-relevant diseases for Hodgkin lymphoma. In general, we observed that the HPO-Orphanet+ graph enriched the existing rare disease knowledge resources and thus be able to provide better diagnostic suggestions. A web-based tool was implemented to visualize diagnostic suggestions and Figure 4 shows an example of this differential diagnostic decision aid interface by considering Hodgkin lymphoma as center node.

In addition, we compared top 15 disease recommendations among the HPO-Orphanet+, eRAM, and Phenomizer using Hodgkin lymphoma as a use case. Since many rare diseases are commonly misdiagnosed as common diseases, it is essential to link common and rare diseases at the early stage of diagnosis to assist in diagnostic decision support. Compared to the Phenomizer and eRAM, we found that the HPO-Orphanet+ graph is more capable of detecting such associations between rare and common diseases. Top 15 diagnostic suggestions for the HPO-Orphanet+, eRAM, and Phenomizer can be found in Table 5.

Text change (page 12, lines 227-237): Table 5 shows top 15 differential diagnostic candidates for Hodgkin lymphoma between the HPO-Orphanet+ graph and two existing diagnostic tools (Phenomizer and eRAM). The HPO-Orphanet+ graph identified 46.7% (7 out of 15) common diseases and 53.3% (8 out of 15) rare diseases. Specifically, chronic obstructive airway disease,
diabetes mellitus, atrial fibrillation, glaucoma, coronary heart disease, degenerative polyarthritis, and chronic kidney insufficiency are common diseases that share the most similar phenotypes with Hodgkin lymphoma, which are considered to be potential candidates for misdiagnosis of Hodgkin lymphoma. While differential diagnostic candidates provided by the Phenomizer are all rare disease. Similarly, the eRAM generates 93.3% (14 out of 15) rare diseases but only 6.7% (1 out of 15) common diseases.

Since many rare diseases are commonly misdiagnosed as common diseases, it is essential to link common and rare diseases at the early time of diagnosis to assist in diagnostic decision support. Compared to the Phenomizer and eRAM, the HPO-Orphanet+ graph is more capable of detecting such associations.

3. Other minor comments are the following ones:

Response 2.3: We appreciate the reviewer’s comments on these minor issues. We have fully addressed them as follows:

3.1 In the introduction (line 45), D3N needs some citation and a brief description.

Response 2.3.1: We thank the reviewer’s comment on this. We no longer used the term D3N and deleted all such term throughout the entire manuscript. Instead, we used the term HPO-Orphanet+ to indicate the enriched graph. Introduction of this new term was added in page 4, lines 70-73. We also added this name in list of abbreviation in page 14 (as shown in Response 2.1).

3.2 (Line 73) Association rules were not proposed in [18], this is a reference to a tool implemented in R (arules). Please cite the former work of R. Agrawal & R. Srikant (Fast algorithms for mining association rules, VLDB 1994).

Response 2.3.2: We thank the reviewer’s comment on this. We have replaced the reference with former work “Fast algorithm for mining association rules” by R. Agrawal and R. Srikant. Detailed update could be found in reference [29] (page 5, line 98).
3.3 (Line 74) Please provide the citations to the original papers that presented these interestingness metrics.

Response 2.3.3: We thank the reviewer’s comment on this. We have added the book “Real and Complex Analysis” written by Walter Rudin as the original references for these metrics. Detailed updates could be found in reference [30] (page 5, line 101).

3.4 (Lines 103 - 104) From my opinion, the D3N is the phenotype-disease bipartite graph not the graph of diseases similarities.

Response 2.3.4: We thank the reviewer’s comment on this. We no longer used the term D3N and deleted all such term throughout the entire manuscript. Instead, we used the term HPO-Orphanet+ to indicate the enriched graph. (text change can be found as shown in Response 2.1).

3.5 Figure 3, please indicate the point in the x-axis at which support is below the given threshold (5E-06). Probably a better method to retain good co-occurrences is to take the elbow in the curve as reference point.

Response 2.3.5: We thank the reviewer’s comment on this. We have made significant updates on the way to select top associations. Specifically, we first use support to choose the frequent itemsets. In order to select the optimal threshold for support, we chose the average support as the minimum support (5E-06). We then used confidence to choose the optimal associations out of those selected frequent itemsets. Similarly, we chose the average confidence as the minimum confidence (0.05). Figure 2 shows the curves do not have much fluctuation with values lower than the threshold. As a result, 13,742 associations (see additional file 1) were identified. The threshold point is marked on the curve. Some discussion on the selection of average value and elbow criterion was also added.

Text change (page 9, lines 177-182): To select the optimal associations, we set average support score 5E-06 as the threshold to first select 31,211 frequent itemsets and we then set average confidence score 0.05 as the minimum confidence to finalize 13,742 associations (see additional file 1). To further validate the selection of thresholds, as shown in Figure 2, we found that both
support and confidence value didn’t have much fluctuation after dropping below their average values (the threshold point is marked on the curve).

Text change (page 13, lines 246-249): In addition, according to some existing studies [46, 47], we also set thresholds as the average of metrics (e.g., support and confidence) to select optimal associations. In the future, we will make an optimal threshold selection scheme combining both average value and elbow criterion [48] in association rule mining.

3.6 (Line 131) "bipartite graphs derived based on" - > "bipartite graphs based on"

Response 2.3.6: We thank the reviewer’s comment on this. We have fixed this. (page 7, line 146)

3.7 (Line 187) "The increment of the average degree for combined graph" This is not true according to Table 3 where EMR graph has a higher average degree than the combined one. Some error?

Response 2.3.7: We thank the reviewer’s comment on this. Since we have updated the way to identify top associations using minimum support and confidence, all consequent results were also updated accordingly, including Table 3.

Table 3 describes graph characterization for bipartite graphs generated from the HPO-Orphanet, EMR, and HPO-Orphanet+. In this updated table, we found that the HPO-Orphanet graph holds the lowest average degree and the HPO-Orphanet+ produces the highest average degree.

Detailed text change and update can be found in page 10 and Table 3.

3.8 In future work it would be interesting to study the extension of Phenomizer with EMR-derived knowledge, and compare if this enrichment improves its performance.

Response 2.3.8: We thank the reviewer’s comment on this. We have added some discussion points on extending our current work with the study conducted by Jiang et al (text change can be found as shown in Response 2.1).

Reviewer 3:

1. Using SemMedDB may have some limitations due to the low precision of semantic predications generated by SemRep. If authors do not have solutions or other alternative resources (such as other rare disease knowledge base) as gold standard, they should mention this as a limitation.

Response 3.1: We appreciate the reviewer’s comments. We have added it in discussion section. The detailed contents are written as shown below.

Text change (page 14, lines 258-262): We used the SemMedDB to measure the IEP of knowledge enrichment. However, some evidences indicated that the SemMedDB is not so accurate due to the limitation of the extraction algorithms used. For example, the SemRep (the generator for SemMedDB) yielded about 75% precision on information extraction [50]. In the future, we will incorporate more disease and phenotype knowledge bases with human annotated associations to measure the knowledge enrichment.

2. The reason why selecting Wilson's disease and Hodgkin lymphoma is not well described.

Response 3.2: We appreciate the reviewer’s comment on this. We have made significant updates on evaluation part and only selected Hodgkin lymphoma as the use case. The reason for choosing Hodgkin lymphoma was added in introduction section.
Text change (page 2, lines 35-39): Symptoms could be treated as phenotypes in symptomatic diagnosis. Taking Hodgkin lymphoma as an example, since symptoms of Hodgkin lymphoma are very similar to other diseases or conditions, such as Cytomegalovirus, Sarcoidosis, and Toxoplasmosis [5], it is meaningful to use underlying disease-phenotype associations to accelerate early differential diagnosis and largely shorten the diagnostic odyssey for patients.

3. In the first line of abstract, "from literature" should be changed to "from the literature"

Response 3.3: We appreciate the reviewer’s comment on this. We have fixed this in page 1, line 9.

4. In the 20th line of page 2: "percentile" -> "the percentile"

Response 3.4: We appreciate the reviewer’s comment on this. We have made significant updates on evaluation and percentile is no longer used in the updated version.

5. On page 7, line 17, "V is the number of vertices in graph" -> "the graph"

Response 3.5: We appreciate the reviewer’s comment on this. We have fixed this in page 8, lines 152 and 157.

Reviewer 4:

1. There is no detail review of the existing literature. The second paragraph in background section serves as a quick review of some related work. However, only a few work was mentioned, and none of them are directly related to enriching phenotype-disease association. In addition, even though there might not be many related work on phenotype-disease association, there are many existing work on gene-disease association, therefore authors should review these


Response 4.1: We appreciate the reviewer’s comments. We have made a thorough literature review on disease-gene association as well as disease-phenotype association. In particular, we have used eRAM, an existing rare disease encyclopedia, to conduct evaluations in this study. Newly added literature review is showed below.

Text change (pages 3-4, lines 46-67): Some other existing studies investigated the mining of associations between diseases and genes. For example, Zhang et al. combined the Latent Dirichlet Allocation (LDA) [9] with network-based computational approach [10] to discover disease-gene associations from large amount of PubMed literature [11]. Piro et al. developed a classification approach to predict disease-gene associations [12]. By leveraging a network distance measure and a random walk algorithm, Kohler et al. presented a method to prioritize candidate genes for hereditary disorders [13]. However, all of these studies focused solely on extracting information from literature or knowledge bases. It often takes substantial time and effort before the suspicion of a rare disease is even raised to utilize those resources due to its rarity. There are some related studies utilizing either electronic medical record (EMR) or
literature or both to investigate diseases, phenotypes and their associations. For example, Xu et al. introduced text mining result of disease-phenotype associations by analyzing sentences from MEDLINE [14]. In another study, Garcelon et al. described a text mining based analysis leveraging TF-IDF to discover associations between clinical phenotypes and rare diseases [15]. Their results showed that phenotypes identified in EMR can be a useful source of evidence to provide rare disease specialists with candidate phenotypes. The eRAM is an encyclopedia of rare disease annotations mined from 10 million scientific publications and EMR [16]. The authors of eRAM implemented a web-based tool to provide clinicians with next-step information of disease-disease associations in addition to disease-phenotype associations. The tool systematically incorporates disease-phenotype associations of rare diseases from both published medical literatures and clinical data. Hassan et al. investigated on extracting associations between rare diseases and phenotypes to enrich existing ontology [17]. The Phenomizer [18] is a clinical diagnostic tool that aims to help clinicians to identify the potential diagnostic candidates. It is built based on the HPO, Orphanet and Online Mendelian Inheritance in Man (OMIM) [19]. Unfortunately, EMR was not incorporated in [17] and [18].

2. I am not an expert on phenotype-disease association, so my question might be too naive. Is it a strong association between phenotype and diseases, particularly some rare diseases? Why it is important to mine phenotype-disease association for rare diseases? Authors should present clearer motivations for taking this approach.

Response 4.2: We appreciate the reviewer's comments. We have highlighted our motivation on mining disease-phenotype association in introduction section. Basically, we wanted to emphasis that it is meaningful to use underlying disease-phenotype associations to accelerate early differential diagnosis and largely shorten the diagnostic odyssey for patients. Detailed updates can be found as shown below.

Text change (page 2, lines 26-39): Rare diseases, although individually rare, collectively affect one in ten Americans. Approximately 7,000 rare diseases exist, with more being discovered each year [1]. Patients with rare diseases face diagnostic delay: 40% of rare disease patients are diagnosed incorrectly before reaching a final diagnosis, of which 25% spend between 5 to 30 years on a chaotic journey through numerous referrals, investigations, and disease evolutions from early symptoms to a confirmatory diagnosis of their disease [2]. Although there are many genetic tests available for delivering precision medicine, how to identify patients who may benefit from those genetic tests is not obvious. Many rare diseases can be misdiagnosed as
common diseases due to their rarity. It often takes substantial clinical time and effort before a rare disease is even a suspected diagnosis [3]. The diagnosis pathway of rare diseases is highly dependent on the associated clinical phenotypes, i.e., the observable characteristics, at the physical, morphologic, or biochemical level, of an individual [4]. Symptoms could be treated as phenotypes in symptomatic diagnosis. Taking Hodgkin lymphoma as an example, since symptoms of Hodgkin lymphoma are very similar to other diseases or conditions, such as Cytomegalovirus, Sarcoidosis, and Toxoplasmosis [5], it is meaningful to use underlying disease-phenotype associations to accelerate early differential diagnosis and largely shorten the diagnostic odyssey for patients.

3. on page 3. this work essentially utilizes the co-occurrence information between a phenotype and a disease for identifying the association. This is a common method in data mining domain, but there usually has a window size of 20 or 50 words to limit the distance between two words when examining their co-occurrence relationship. In this paper, there is no window size, but the whole clinical note is used. I do not know how long a typical clinical note is, but if it is like a normal document, it would be a too large context for identifying co-occurrence relationship based on data mining studies. Therefore, the authors should make it clearer on this aspect, and probably discuss the motivations of using the whole note rather than a window size of clinical note for obtaining co-occurrence relationship.

Response 4.3: We appreciate the reviewer’s valuable suggestion. We have added some points in the discussion section to explain this. Detailed updates are showed below.

Text change (pages 13-14, lines 251-257): In this study, we extracted the co-occurrence information between a phenotype and a disease from diagnosis section contained in clinical notes. Specifically, we first split the entire notes into sentences and then applied the aforementioned annotation pipeline on each sentence. In addition, problems in those documents are generally itemized entries as either phrases (e.g., Allergic rhinitis/vasomotor rhinitis) or short sentences (e.g, Her asthma appeared to be very mild), therefore, we didn’t use window size to limit the distance between phenotype and disease. In the future, to generalize the association mining on larger size of documents, we will seek to investigate the selection of appropriate window size for a better performance [49].

4. page 6. DCG is a measure designed for modeling relevance scores at multiple grade level (i.e., more than two). If there is just binary relevance, it is actually common to use average precision
(AP) to model the quality of the ranked list. AP has the advantage of having maximum value at 1, whereas DCG can increase indefinitely. In addition, current paper does not present the meaning of p in DCG clear enough. Maybe an example to illustrate the idea is better.

Response 4.4: We appreciate the reviewer’s comments. We no longer used DCG to measure the ranking since this metric cannot well evaluate knowledge enrichment.

The aim of this study is to validate the effectiveness of enriched knowledge base versus the original one. First of all, IEP was used to quantify the enrichment of phenotype-disease associations compared to the SemMedDB. In addition, in the evaluation, we added a new experiment to compare three bipartite graphs: HPO-Orphanet graph, EMR graph, and Enriched HPO-Orphanet graph (HPO-Orphanet+).

5. page 11. The Phenomizer is a generic clinical diagnostic tool, so it can be a baseline, but it should not be the only baseline for comparison. The authors should select from the related work one or several recently published rare disease diagnostic tools to use as baselines. Only through this, it makes sense to claim the innovation of the proposed method.

Response 4.5: We appreciate the reviewer’s comments. We made significant updates on the evaluation. Following the reviewer’s suggestion, we introduced a recent published study named eRAM in evaluation. Specifically, the eRAM is a web-based tool to provide clinician with next-step information of disease-disease associations in addition to disease-phenotype associations. We used the eRAM to generate gold standard for this evaluation. Since data generated by the eRAM is coming from both literature and EMR, it is unbiased and fair to compare the eRAM with our tool. Detailed illustration on how to generate the eRAM gold standard can be found in pages 8, lines 161-168.

Text change (page 8, lines 161-168): To prepare the experiment, for any disease to be tested, we used the three aforementioned graphs to rank suggested diseases with descending order of Jaccard similarity score. We combined two disease-phenotype association files namely “eRAM Integrated Phenotype.txt” and “eRAM Integrated Symptom.txt” provided by the eRAM. Based on 5,356 curated diseases and their associated phenotypes/symptoms obtained from the files, we calculated Jaccard similarity score between each pair of diseases using Equation 4. Such disease-disease similarity was used as a gold standard on differential diagnosis. In this evaluation, we
first validate diagnostic candidates generated by the three bipartite graphs using the gold standard. We then compared the top 15 differential diagnostic candidates generated by the HPO-Orphanet+ graph, Phenomizer, and eRAM.

The aim of this study is to validate the effeteness of enriched knowledge base versus the original one, therefore, we compared performances on differential diagnostic suggestions for three bipartite graphs leveraging the gold standard provided by the aforementioned eRAM: original HPO-Orphanet Graph, EMR Graph, and Enriched HPO-Orphanet Graph (the new name is HPO-Orphanet+). We observed that the HPO-Orphanet+ graph shows the highest sensitivity for detecting the right similar diseases according to the gold standard, while using the original HPO-Orphanet based graph yields the lowest sensitivity (As shown in Figure 3). Relevant descriptions of this evaluation can be found in section “Rare disease differential diagnostic suggestions – use case study” in page 12.

Text change (page 12, lines 218-226): Sensitivity and specificity for generating differential diagnostic suggestions for Hodgkin lymphoma with different graphs is shown in Figure 3. The HPO-Orphanet+ graph shows the highest sensitivity for detecting the right similar diseases according to the eRAM gold standard, while using the HPO-Orphanet graph yields the lowest sensitivity. In addition, specificity does not show significant differences among three graphs, indicating that all of them have similar performances on rejecting non-relevant diseases for Hodgkin lymphoma. In general, we observed that the HPO-Orphanet+ graph enriched the existing rare disease knowledge resources and thus be able to provide better diagnostic suggestions. A web-based tool was implemented to visualize diagnostic suggestions and Figure 4 shows an example of this differential diagnostic decision aid interface by considering Hodgkin lymphoma as center node.

In addition, we compared top 15 disease recommendations among the HPO-Orphanet+, eRAM, and Phenomizer using Hodgkin lymphoma. Since many rare diseases are commonly misdiagnosed as common diseases, it is essential to link common and rare diseases at the early stage of diagnosis to assist in diagnostic decision support. Compared to the Phenomizer and eRAM, we found that the HPO-Orphanet+ graph is more capable of detecting such associations between rare and common diseases. Top 15 diagnostic suggestions for the HPO-Orphanet+, eRAM, and Phenomizer can be found in Table 5.
Table 5 shows top 15 differential diagnostic candidates for Hodgkin lymphoma between the HPO-Orphanet+ graph and two existing diagnostic tools (Phenomizer and eRAM). The HPO-Orphanet+ graph identified 46.7% (7 out of 15) common diseases and 53.3% (8 out of 15) rare diseases. Specifically, chronic obstructive airway disease, diabetes mellitus, atrial fibrillation, glaucoma, coronary heart disease, degenerative polyarthritis, and chronic kidney insufficiency are common diseases that share the most similar phenotypes with Hodgkin lymphoma, which are considered to be potential candidates for misdiagnosis of Hodgkin lymphoma. While differential diagnostic candidates provided by the Phenomizer are all rare disease. Similarly, the eRAM generates 93.3% (14 out of 15) rare diseases but only 6.7% (1 out of 15) common diseases. Since many rare diseases are commonly misdiagnosed as common diseases, it is essential to link common and rare diseases at the early time of diagnosis to assist in diagnostic decision support. Compared to the Phenomizer and eRAM, the HPO-Orphanet+ graph is more capable of detecting such associations.