BIOINFORMATICS

Vol. 00 no. 00 2015
Pages 1-7

Decision support for the diagnosis of rare diseases based on symptoms

Marie-Sophie Friedl [1,2,3], Carolin Prexler [1,2,3], Maria Schelling [1,2,3], Gabi Kastenmüller [3], Andreas Ruepp [3] and Jan-Dominik Quell [3]

[1]Ludwig-Maximilians Universität München, Germany
[2]Technische Universität München, Germany
[3]Helmholtz Center Munich - German Research Center for Environmental Health (GmbH), Institute of Bioinformatics and Systems Biology, Neuherberg, Germany

ABSTRACT

Motivation: There are more than 6 000 rare diseases, each of which has a prevalence of less than 5 cases per 10 000 people. It is highly unlikely that a general practitioner will ever treat more than one patient suffering from a particular rare disease. Databases about the phenotypes of rare diseases and differential diagnostic algorithms like Phenomizer are valuable tools that help those doctors to yet correctly diagnose a rare disease. To guarantee a good performance of a diagnostic algorithm, an informative database has to be provided and the algorithm has to be accurately validated on real patient data.

Method: We present an implementation of the Phenomizer diagnostic algorithm that works on the manually curated PhenoDis database. Our tool compares a set of observed symptoms against all diseases annotated in PhenoDis using an built-in ontology of symptoms, and visualises the most similar diseases within a network of all diseases. The validation of Phenomizer for PhenoDis on real patient data from the literature proves that the tool is able to make a valid diagnosis in more than 60% of all cases. Extending the original algorithm with features for weighting symptoms according to their frequency even increases the sensitivity of the method.

Availability: PhenoDis for Phenomizer is available on GitHub at https://github.com/marie-sophie/mapra.

Contact: M.friedl@campus.lmu.de,
carolin.prexler@campus.lmu.de,
maria.schelling@campus.lmu.de

1 INTRODUCTION

Recent surveys from health care organisations have shown that patients affected by rare diseases often face diagnostic delays or even have to deal with incorrect or inaccurate diagnoses [29][33]. The diagnostics is often challenging because the level of expertise varies between clinicians and because individual patients with a given disease may have different, partially overlapping combinations of symptoms due to pleiotropy and variable expression [21]. The diagnostic process is further complicated by the large number of Mendelian and chromosomal disorders including more than 6 000 rare disorders. Additionally, their clinical features are often shared among many diseases.

The application of a differential diagnostic tool can overcome these problems and largely improve and fasten the process of diagnosing rare diseases. Genetic databases and ontologies, such as the Human Phenotype Ontology (HPO) [28] and the Online Mendelian Inheritance in Man (OMIM) [12], provide new possibilities to design diagnostic algorithms. These algorithms attempt to identify candidate diseases which best explain a set of clinical features. But varying degrees of specificity of symptom terms and the presence of features unrelated to the disease itself complicate the workflow of the differential diagnostic process. When developing a differential diagnostic algorithm, it is important to deal with these complications. There are already some algorithms available, e.g. DXPlain and Isabel [6].

Phenomizer [21] is another of these differential diagnostic algorithms. It is based on the structure of HPO, accesses information from OMIM and was validated using 4 400 simulated patients. In this publication, we present an implementation of the Phenomizer algorithm using data of a different database, namely PhenoDis. PhenoDis combines descriptions of diseases from diverse resources, and expands, substantiates and standardises this information by manual annotation. We validate the results of our implementation of Phenomizer using entries of OMIM and real patient data from the literature. With this approach, we provide a more realistic validation than the original publication. Furthermore, we integrate different refinements to improve the results of Phenomizer. With our work, we make a diagnostic algorithm available to users of the PhenoDis database and demonstrate that Phenomizer performs especially well on rare diseases and is able to deal with complex case descriptions.

2 MATERIALS AND METHODS

2.1 Konstanz Information Miner

KNIME is an open-source analytics and modelling platform which is based on Java inclosing an integrated development environment (IDE) and an extensible plug-in system [4][20]. KNIME provides diverse nodes computing specific isolated tasks. These nodes can be combined individually by the user to build a workflow, for example to load, analyse, process and visualise data of every kind. It is also possible to expand or adjust an already existing workflow suiting your objective.

We use KNIME to develop and evaluate a differential diagnosis workflow. The core component of our differential diagnosis pipeline is the integration

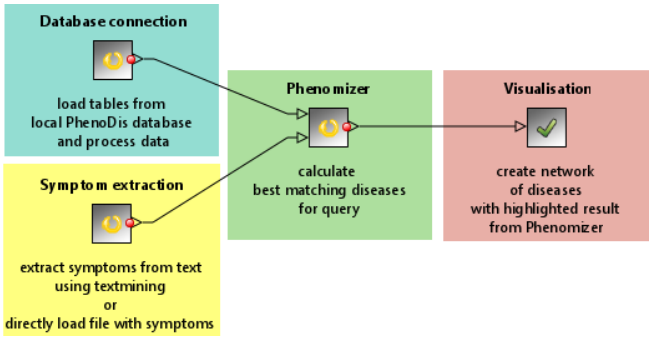

Fig. 1. Overview of the main steps of our workflow: the connection to the local PhenoDis database including data processing (blue), the generation of the query by either using text mining or directly loading a file of symptom names (yellow), the actual Phenomizer algorithm computing the best matching diseases for the query (green) and the visualisation of the top scoring results in a network of diseases (red).

of Phenomizer as a new node (Fig. 1, node "Phenomizer"). The built-in text-processing and plotting nodes of KNIME are applied to evaluate and to visualise the results of Phenomizer. For our whole KNIME workflow including processing of data, the Phenomizer algorithm, text mining and validation, see the supplementary material, figure S1.

2.2 PhenoDis database

The PhenoDis database used during this project is a comprehensive database collecting information about diseases, their phenotype and their genetic background including prevalence and age of onset. This database includes annotated information about diseases and symptoms from OMIM [12], Orphanet [27] and HPO [28], and is manually curated by Helmholtz Center Munich. For the implementation of Phenomizer, we use information about diseases, symptoms and their associations with each other.

The description of the symptoms in PhenoDis is based on the vocabulary and the ontology of HPO. If appropriate, terms of the MedDRA vocabulary [24] or OMIM are used in addition. In general, an ontology is a data model to represent relationships between different entities with defined attributes [28]. Our ontology is a directed acyclic graph that models hierarchical parent-child relationships between symptoms.

During the process of annotation, symptoms are assigned to the diseases they are associated with. The annotations of diseases with symptoms are realised using the relationships described in the ontology. For approximately half of these annotations, information about the frequency of the symptoms is available. Some symptoms are more often observed in patients suffering from a certain disease than others. According to Orphanet, there are three possible frequency terms to describe how prevalently a symptom occurs with a disease: occasional, frequent and very frequent. PhenoDis uses the Orphanet terminology. Other frequency descriptors in the database stem from external HPO annotation. If available, annotators from the Helmholtz Center also provide absolute frequency values.

2.3 Preprocessing of data

We had to preprocess the data from PhenoDis to ensure consistency (Fig. 1, node "Database connection") . As PhenoDis integrates data from several sources, the hierarchy of the symptoms may contain directed cycles. Filtering out all symptoms that are not linked to HPO creates an acyclic hierarchical ordering Phenomizer can work with. Terms in the other used data tables were also filtered out if they were not from HPO. Diseases without any symptoms were removed because these diseases are not annotated yet and are not considered during the calculations done by Phenomizer. Overall, our final dataset consists of 56 184 symptoms, 7 554 diseases and 186 786 associations between these terms.

The last preprocessing step deals with the frequency annotation of the disease-symptom associations. To use a uniform vocabulary, we map frequency terms from HPO and absolute frequency values to the Orphanet terms. For detailed information about the mapping, see supplementary material, table S1. Disease-symptom associations with unknown frequency are also mapped to the Orphanet term "frequent". It is important to generate a uniform vocabulary because the frequency terms are used for an extension of the Phenomizer algorithm.

2.4 Implementation and extension of Phenomizer

Original algorithm Phenomizer is a diagnostic algorithm which uses semantic similarity metrics to compare queries of symptoms with diseases in a given database [21]. The goal is to determine diseases best matching a given set of query symptoms.

The basic idea of Phenomizer is to consider the distance between symptoms in the ontology of the PhenoDis database. The distance is then used to identify diseases with symptoms similar to a query. The similarity between two symptoms depends on the specificity of the common ancestors in the hierarchy. The specificity of a symptom in the ontology is reflected by its information content (IC). The IC is defined as $-\log p(t)$ where $p(t)$ gives the proportion of diseases annotated to a symptom term t or any of its descendants. The IC of a symptom increases with increasing distance to the root.

Then, the similarity between two symptom terms t_1 and t_2 is calculated as the IC of the most informative common ancestor as shown in equation 1. The expression $A(t_1, t_2)$ denotes the set of all common ancestors of t_1 and t_2.

$$\mathrm{sim}(t_1, t_2) = \max_{a \in A(t_1, t_2)} -\log p(a) \tag{1}$$

Equation 1 yields a pairwise similarity measure for comparing two symptoms with each other. This pairwise similarity is used to determine a similarity score between a given set of query symptoms Q and a set of symptoms D of any disease in the database. The similarity score is calculated by identifying the best match for each symptom in the query among the symptoms of the disease and taking the average over all pairwise similarity values (Eq. 2).

$$\mathrm{sim}(Q \to D) = \frac{1}{|Q|} \left(\sum_{t_1 \in Q} \max_{t_2 \in D} \mathrm{sim}(t_1, t_2) \right) \tag{2}$$

Equation 2 does not take into account that there might be symptoms in D that are not assigned to any symptom in Q. To avoid this issue, a symmetric similarity measure is introduced as shown in equation 3.

$$\mathrm{sim}_{\mathrm{symmetric}}(Q, D) = \frac{\mathrm{sim}(Q \to D) + \mathrm{sim}(D \to Q)}{2} \tag{3}$$

Finally, the diseases leading to the highest symmetric similarity scores are considered as a possible diagnosis for the set of query symptoms.

This symmetric similarity score depends on several factors including the number and the specificity of the symptoms in the annotation of the diseases and in the query. So, a general statement which score represents a good match cannot be made. To overcome this problem, a p value is calculated for each search result. The p value "indicates the probability of obtaining the same or higher similarity scores by a randomly generated query set of the same size" [21].

We estimate the p values by precalculating an empirical score distribution for each disease in the database. For each disease entry, we generated 100 000 random queries of lengths one to ten and stored the resulting similarity scores on disk. The score distributions for queries with more than ten symptoms are approximated by using the distributions for queries of length ten.

The p value of a certain result is the fraction of scores from the score distribution greater than or equal to the observed similarity score. Additionally,

the obtained p value was adjusted using the Benjamini-Hochberg procedure for multiple testing correction [3]. If an observed similarity score is only rarely obtained by chance, and therefore has a small p value, it is considered statistically significant.

Phenomizer with weighting When trying to make a diagnosis, symptoms which are (very) frequently observed with a disease should be taken into account more strongly than symptoms occurring rarely with the disease. We realise that by extending equation 2 of the original Phenomizer algorithm. Each pairwise similarity is multiplied by a weight reflecting the frequency of the symptom of the considered disease leading to the definition of the weighted similarity. Equations 4 and 5 show the similarity of a query to a disease and of a disease to a query, respectively.

$$\mathrm{sim_w}(Q \to D) = \frac{1}{|Q|}\left(\sum_{t_1 \in Q} \mathrm{weight}(D, t_2) \max_{t_2 \in D} \mathrm{sim}(t_1, t_2)\right) \quad (4)$$

$$\mathrm{sim_w}(D \to Q) = \frac{1}{|D|}\left(\sum_{t_1 \in D} \mathrm{weight}(D, t_1) \max_{t_2 \in Q} \mathrm{sim}(t_1, t_2)\right) \quad (5)$$

A diagnosis based on the weighted similarity scores in equations 4 and 5 takes not only the similarity of the symptoms, but also the frequency of the symptoms into account. There are two possible ways to realise the weighted similarity score. First of all, it can be calculated as a double-sided weighted score by taking the average over the two pairwise similarity scores resulting from equations 4 and 5 as shown in equation 6.

$$\mathrm{sim_{double\text{-}sided}}(Q, D) = \frac{\mathrm{sim_w}(Q \to D) + \mathrm{sim_w}(D \to Q)}{2} \quad (6)$$

But it is also possible to only take the weight into account when calculating the similarity of the query to a disease as given in equation 4 while using the unweighted pairwise similarity score from equation 1 to describe the similarity of a disease to the query. This leads to the definition of the one-sided weighted similartiy score in equation 7.

$$\mathrm{sim_{one\text{-}sided}}(Q, D) = \frac{\mathrm{sim_w}(Q \to D) + \mathrm{sim}(D \to Q)}{2} \quad (7)$$

KNIME node We implemented the Phenomizer algorithm as a KNIME node (Fig. 1, node "Phenomizer") which takes the information about the symptoms, the diseases, the association between both and the ontology from PhenoDis as input. Furthermore, a query of symptoms has to be provided. This query can consist of symptom IDs referring to PhenoDis entries or of HPO terms. The node returns a similarity score and an optional p value for each disease in PhenoDis as a sorted table (Fig. S3). Additionally, the user can choose between different variants of the Phenomizer algorithm and adjust the size of the output (Fig. S2).

2.5 Visualisation

We provide a second KNIME node called PhenomizerToNetwork to visually examine the results of Phenomizer (Fig. 1, node "Visualisation"). This node is a wrapper for Cytoscape (version 3.2.1) and is based on the KNIME nodes from [26]. PhenomizerToNetwork requires two inputs: the results from Phenomizer and a matrix with similarity scores for all pairs of diseases in PhenoDis. Using this data, the node creates a network of all diseases. The resulting network contains a vertex for each disease and edges connecting diseases with a high similarity. The vertices corresponding to the top scoring diseases from the Phenomizer results are highlighted. The integration of Cytoscape into the node is required to display this network.

2.6 Text mining and validation

Used data To validate the results of our implementation, we use two kinds of data. First of all, we use the diseases and their clinical synopses from the OMIM database. The complete OMIM database was downloaded and filtered for all diseases which are also included in PhenoDis and have at least one annotated symptom, leaving 5 760 test cases. These cases are used to compare the different versions of Phenomizer.

We also apply some case studies in order to simulate a realistic diagnostic problem. This second approach deals with 21 articles of 31 real cases describing eight different rare diseases (Acromelanosis [1][11][18][30], Blue Diaper Syndrome [10][23], CADASIL [2][8][32], Hawkinsinuria [5][7][25][31], Leigh Syndrome [9][14][22], Congenital Muscular Dystrophy [13][15][16], Neonatal Hemochromatosis [19] and Systemic Lupus [17]). To analyse these articles, we apply four approaches for extraction of symptoms with varying levels of detail (Fig. 1, node "Symptom extraction"):

1. Usage of the whole article: Extraction of all symptoms mentioned in the text.
2. Usage of case descriptions: Only consideration of the sections describing the actual case; publications including more than one case description are split and each case is analysed separately.
3. Usage of the abstract: Manual extraction of symptoms from the abstract.
4. Manual analysis of the whole article.

Text mining Text mining is used to extract keywords, which map to known terms in HPO, generating a query of symptoms. This query serves as input for Phenomizer. The whole workflow for the text mining procedure is realized in KNIME and is applied to all clinical synopses in OMIM, the whole articles of the case studies and the individual case descriptions.

Calculated quantities For the data described above, we calculate two quantities to measure the accuracy of Phenomizer. First of all, we determine the sensitivity of the Phenomizer algorithm. In our case, the sensitivity is defined as the fraction of results of Phenomizer where the correct diagnosis is included in the best suggested diseases. This provides a general measurement of the accuracy of Phenomizer.

Additionally, the rank of the correct diagnosis is considered, following the idea from the original publication [21]. In case of ties where two or more diseases have the same score, the average rank is returned. Using this quantity, low ranks imply a high accuracy of Phenomizer.

3 RESULTS

3.1 Network of similar diseases

We use the Phenomizer algorithm to compare all 7 554 diseases in our database to each other. To do so, similarity scores (Eq. 3) for all pairs of diseases are calculated. The resulting similarity scores lie between zero and ten. The resulting matrix can serve as a basis for generating a network of diseases with the PhenomizerToNetwork node. The structure of this network depends on the choice of the connections. Two diseases are connected by an edge if their similarity score lies above a certain predefined threshold. In particular, this cutoff influences the number and size of the connected components within the network. The number of components increases with increasing threshold while the average size of the connected components and the average degree of the nodes decrease. A threshold of 4.0 for the similarity is reasonable as the resulting network consists of connected components of appropriate size. A systematic validation to determine a threshold for the score is currently infeasible as a gold standard for the similarity between diseases is not available.

Analysing the connected components of the network indicates that the nodes within a component represent related diseases. So, a component can be interpreted as a cluster of biologically similar diseases annotated with similar symptoms. For example, the component shown in figure 2 consists of six diseases which are all

associated with an unusal iron level (hemochromatosis or anemia). The corresponding network consists of 311 connected components excluding singleton nodes and an average degree of 4.6. This network can also be used to visualise the results from Phenomizer. The top scoring diseases are coloured differently in the network. This visualisation allows to analyse whether the results from Phenomizer lie in the same connected component, i.e. a family of related diseases. For an exemplary visualisation, see supplementary material, figures S4 and S5.

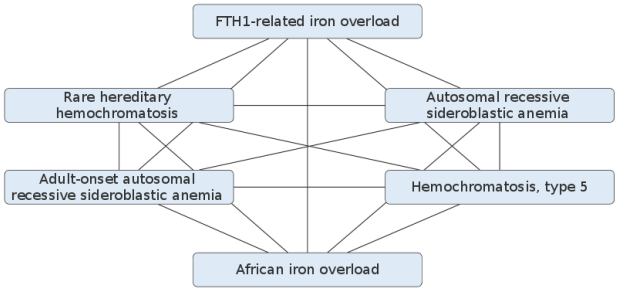

Fig. 2. Exemplary connected component of related diseases with six nodes from the network of 7 554 diseases from PhenoDis. Edges were drawn between diseases with a similarity score > 4.0

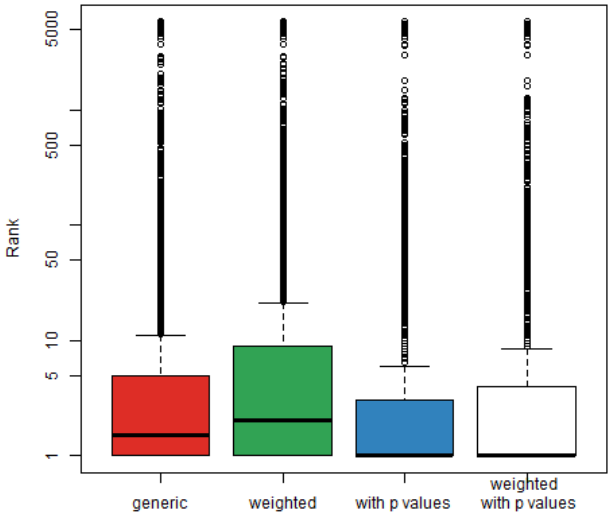

Fig. 3. Rankings of correct diagnosis of the clinical synopses from OMIM by the four versions of Phenomizer: generic, with p values, weighted and weighted with p values

3.2 Validation of Phenomizer

To compare the accuracy of different versions of Phenomizer, we run the algorithm on the clinical synopses extracted from OMIM and record the resulting rankings of the correct diagnosis. The rank of the correct diagnosis is its position in the ordered output. Examining the rankings provides a detailed evaluation of Phenomizer, especially for the comparison of the four different versions:

1. Generic Phenomizer (based on Eq. 2)
2. Weighted Phenomizer (based on Eq. 7)
3. Phenomizer with p values (see subsection 2.4, Phenomizer with p values)
4. Weighted Phenomizer with p values

When analysing the results, we can demonstrate that the overall performance of the generic version is very good (Fig. 3): around 50% of the ranks are below 5 and only 5% of the diseases obtain a rank above 10. However, the accuracy of Phenomizer always improves when using p values regardless whether the weighted or the unweighted similarity score is calculated. For the unweighted Phenomizer with p values, almost 95% of the diseases from OMIM have a rank below 5.

The results for the weighted Phenomizer version (Fig. 3) are based on the one-sided weighted similarity (Eq. 7). The weighted Phenomizer takes the frequency with which symptoms are associated with a disease into account. The one-sided score only uses the frequency when calculating the similarity of the query to a disease (Eq. 4) while the double-sided one also applies the weighted score for the similarity of a disease to the query (Eq. 5). The comparison of the one-sided and the double-sided weighted similarity scores shows a higher performance of Phenomizer for the one-sided one (Fig. S6). We choose three different weighting factors for the three different terms from Orphanet for the frequency of an association between a symptom and a disease: 0.5 for occasional, 1 for frequent and 1.5 for very frequent. The ranks for the weighted Phenomizer are slightly higher than the ones for the generic version.

For general practitioners, it is especially difficult to diagnose a rare disease correctly. A differential diagnostic algorithm like Phenomizer should not only perform well in general, but especially also for rare diseases to provide high quality decision support during the diagnostic process. Therefore, we also compare the performance of Phenomizer with p values for common and rare diseases. A disease from OMIM is considered as rare if it is also annotated in Orphanet. Examining the ranks for common and rare diseases, it becomes obvious that Phenomizer does not only perform equally well for rare diseases, but even better than for common diseases (Fig. 4).

In conclusion, the validation with OMIM shows that Phenomizer performs well, especially when using p values while the application of weights does not lead to an improvement of the results. Phenomizer is also able to make a valid suggestion of a diagnosis even for rare diseases.

Additionally, we also use known case studies to validate Phenomzier on real patient data. Here, we compare the sensitivity values for three different versions of symptom extraction (whole article, case description, abstract) instead of recording the ranks. The sensitivity is the fraction of cases where the correct diagnosis lies in the top scoring diseases. Since it is mainly important that the correct disease is listed highly in the results, it is plausible to use the sensitivity, especially for real patient data. For each approach of extraction, the unweighted and weighted Phenomizer with p values are analysed. The results for this second validation are shown in figure 5. The sensitivity is given as the fraction of correct diagnoses with a rank ≤ 30, ≤ 20 and ≤ 10. Using the whole article to generate a query

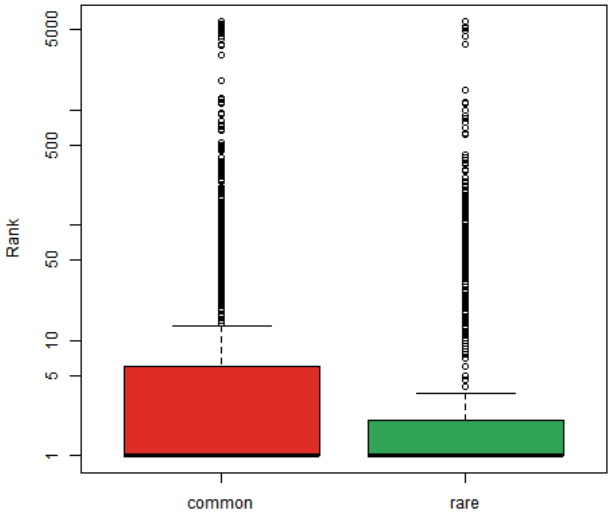

Fig. 4. Rankings of common versus rare diseases from OMIM for the unweighted Phenomizer with p values

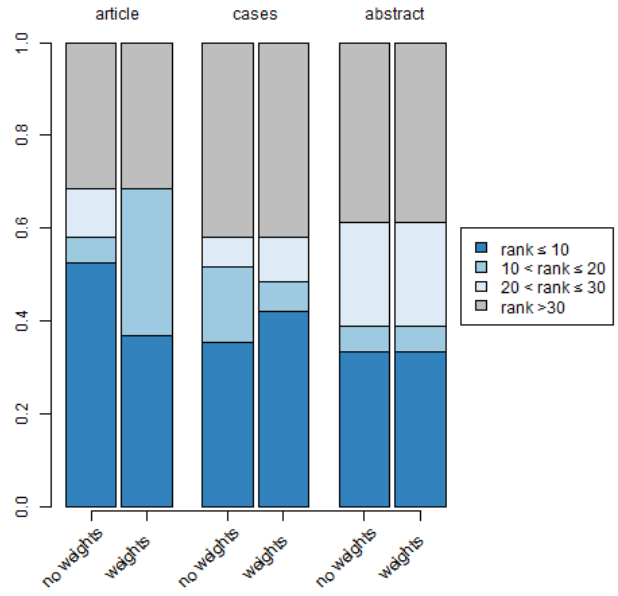

Fig. 5. Fraction of correct diagnoses in the 10, 20 and 30 top scoring diseases for real cases. The performance of the unweighted and weighted Phenomizer are compared for three different methods of symptom extraction: the usage of the whole article, of the case descriptions and of the abstract.

leads to an overall sensitivity of almost 70%, which is the best value compared to the usage of abstracts or of case descriptions. For all three kinds of symptom extraction, the results show that weighted and unweighted Phenomizer predict the same amount of correct diagnoses in the 30 top scoring diseases. So, the overall performance of the two versions is equally well. Differences in the accuracy of the versions only become obvious when examining the more detailed sensitivity values for the 10 and 20 top scoring diseases. The application of the weighted Phenomizer can especially improve the results for the usage of the case descriptions leading to a higher percentage of correct diagnoses in the 10 top scoring diseases.

Summing up, the application of case studies to validate Phenomizer demonstrates that the algorithm is also able to make good predictions of a possible diagnosis for real patient data. In contrast to the validation with OMIM, the weighted Phenomizer achieves equally good results for the case studies as the unweighted Phenomizer.

For the complete rankings of all case studies for the unweighted and weighted Phenomizer with p values, see supplementary material, table S2.

4 DISCUSSION

With our implementation of Phenomizer, we provide a high-performing diagnostic algorithm for the manually curated database PhenoDis. The application of weights supplies an enhancement of the original algorithm. The weights represent the frequency with which the symptoms are observed with a disease. Furthermore, this is also the first time that Phenomizer is systematically validated on real patient data. Additionally, we perform a large-scale validation with data from OMIM.

The validation of Phenomizer with the clinical synopses from OMIM shows an overall high performance of Phenomizer. Synopses obtaining a low similarity score to the respective disease in PhenoDis can be explained by short or uninformative clinical synopses in OMIM. For example, the OMIM entry of “Osteoma of Middle Ear” (259650) contains only two symptoms: “Middle ear osteoma” and “Autosomal recessive”. Such descriptions lead to short and unspecific queries. If the number of symptoms annotated to a disease in PhenoDis is much larger than the number of symptoms in the query, the possibility to obtain the observed similarity score by chance increases, and the p value and the rank for the corresponding query get higher. Therefore, the length and the content of a query should always be taken into account when interpreting the results of Phenomizer.

We also demonstrated that Phenomizer performs better for rare diseases than for common ones. The number of symptoms annotated to a disease in PhenoDis explains the differences in the performance of the algorithm for common and rare diseases. We observe a correlation between the number of annotated symptoms and the accuracy of the diagnostic prediction (Fig. S7). The larger the number of symptoms for a disease, the higher is the chance for a correct diagnosis. Rare diseases are in fact annotated with more symptoms than common diseases in PhenoDis (Fig. S8). This can be reasoned by the composition of PhenoDis as this database focuses on the annotation of rare diseases and integrates data from Orphanet in addition to OMIM. The comparison of the performance for common and rare diseases proves that the Phenomizer algorithm for PhenoDis achieves the aim to support the diagnostic process for rare diseases.

If a disease is annotated with some very frequent symptoms and a patient presents one of these, it is reasonable to assign a higher pairwise similarity score than for an observed rare symptom of this disease. This fact is realised by using weighted scores. We use one-sided weighting (Eq. 7) which is an improvement in contrast to double-sided weighting (Eq. 6) (Fig. S6). We focus on the presence

of very frequent symptoms in a query. Therefore, we only adjust the similarity score of a query to the disease ($\text{sim}(Q \rightarrow D)$) and not the score of a disease to a query ($\text{sim}(D \rightarrow Q)$). The application of the one-sided weighted similarity score takes the different frequencies of symptoms into account, and especially assigns high pairwise similarity scores to pairs of matching, very frequent symptoms.

In addition to the validation with OMIM, we apply a new kind of validation using case studies from the literature. This provides a more realistic validation in comparison to the usage of OMIM or of simulated patients as used in the original publication [21]. Clinical synopses and simulated patient data do not correspond to real patients. Both kinds of data consist of symptom terms exactly matching HPO terms. Besides, real patients show only a certain amount of symptoms while a clinical synopsis is a description of the disease as detailed as possible.

Consequently, case studies are a more realistic, but also more challenging diagnostic task. Symptoms described in a case study can be described on a different level of specificity than the corresponding symptoms annotated with the disease, e.g. the case report in [16] uses the term "weakness" whereas the corresponding OMIM entry (ID 607855) lists the symptoms "Respiratory muscle weakness" and "Muscle weakness, severe, axial and proximal predominance". Besides, case studies often mention symptoms which are not related to the disease, or miss symptoms annotated to the disease.

The articles vary in the level of detail and the number of mentioned symptoms. Many publications have additional sections discussing general rules to diagnose the presented rare disease or describing in which ways the diagnosed rare disease can be distinguished from other more common diseases. On the other hand, some texts are very short and contain hardly any symptoms. For example, the case description of Blue Diaper Syndrome in [23] yields only one reasonable symptom "Discolored, acholic stools" whereas Blue Diaper Syndrome is characterized by 27 symptoms in PhenoDis.

In conclusion, suggesting a diagnosis for a case study is difficult. But even for this challenging input, Phenomizer achieves high sensitivity values (Fig. 5). We observe an improvement of the performance of Phenomizer for some case studies when focusing on the actual case description. This effect is clearly visible for the article about Acromelanosis [1] which has an extremely long discussion section. The number of extracted symptoms is reduced by about 70% and the rank of Acromelanosis improves from 18 to 1.

Another challenge for the diagnosis based on case studies is the generation of a query out of the article describing the patient. This is mainly due to the inherent limitations of the KNIME text mining nodes. Text mining is not able to recognise negated statements like "She was negative for rheumatoid factor" [17]. In addition, text mining has difficulties matching related terms, e.g. "the infant appeared to be deaf" [22] is not identified as a symptom called "deafness".

Despite these general difficulties, the Phenomizer algorithm performs well on the queries generated by text mining. We also extracted the symptoms manually from the abstracts or from the whole articles to overcome these general problems of text mining. These symptoms can be matched more precisely to the corresponding symptom terms in PhenoDis. This leads to an improvement of the ranks, e.g. for Blue Diaper Syndrome [23] (Tab. S2). Hence, the usage of a more sophisticated text mining tool can even further improve the results of Phenomizer.

The analysis of the case studies shows small differences in the sensitivities for the unweigthed and weighted Phenomizer with p values (Fig. 5). These differences are not entirely representative because we consider only 21 publications about eight rare diseases. But the results do not show a decline in the performance when using the Phenomizer algorithm with weights as it is observed for the OMIM data. The application of weights is based on the assumption that (very) frequent symptoms are more often observed in a query than rare symptoms. But this assumption does not hold for the clinical synopses in OMIM. It is unlikely that symptoms unrelated to the disease are annotated in the corresponding OMIM entry. Hence, (very) frequent symptoms are present in the query as well as rare symptoms. The case studies provide a more valid validation of the weighted similarity score. For some case studies, the usage of weights even improves the ranks.

With these observations, we show that the application of weights has the potential to improve the performance of Phenomizer. Consequently, the Phenomizer version with weights and p values can be helpful when trying to diagnose real patients.

The visualisation of the results from Phenomizer makes it possible to examine the similarity of the top scoring diseases. If these diseases cluster in the network, they are from the same family of diseases. This observation gives a clue which family the correct diagnosis could belong to. For an example, see supplementary material, figure S9. It is likely that the patient is affected by a disease from the same family as the predicted diseases. So, this visualisation helps to interpret the results of Phenomizer and supports the diagnostic process.

The PhenoDis database is still under development. As it is manually curated, the database is not completely annotated at the moment. For example, the information about the frequency of the symptoms for a certain disease is only provided for about every second entry. The information about the prevalence of a disease is even more sparse as it is given for just a quarter of all diseases in PhenoDis. In case of the frequency data, we also have to deal with non-uniform specifications. It is problematic to treat terms like "1 of 4 cases" the same as "25%" as the first term definitely lacks expressiveness. There are also contradictory descriptions coming from different sources. The incompleteness of the database complicates the validation of Phenomizer. As soon as PhenoDis is fully annotated, Phenomizer has to be re-evaluated and the results are likely to improve further.

5 CONCLUSION

In this project, we provide a KNIME implementation of the Phenomizer algorithm for the PhenoDis database. The interpretation of the results of Phenomizer is facilitated by another tool called PhenomizerToNetwork. It displays the top scoring predictions of Phenomizer within a network of diseases giving information about groups of very similar diseases in the output of Phenomizer.

The validation of Phenomizer with various data from OMIM and from the literature shows an overall good performance with a sensitivity of more than 60%. We demonstrate that Phenomizer makes good predictions for common as well as rare diseases and that its performance on rare diseases is even better than on common diseases. We also implement and compare different extensions of the Phenomizer algorithm. The first extension is the calculation of p values to judge the significance of the similarity scores. The usage of p values leads to a considerable boost in the predictive accuracy of Phenomizer. The second extension introduces weights to distinguish between symptoms which are frequently or rarely associated with a disease. This feature further improves the diagnostic suggestions of Phenomizer when facing real patient data.

The predictive performance of Phenomizer will even increase in the future with the proceeding completion of the PhenoDis database. For example, considering the age of onset during the computational process can be helpful and give another clue to find the correct diagnosis. Another future task will be the integration of the Phenomizer tool into an interactive user interface making Phenomizer available to users unfamiliar with KNIME. After this is realised, Phenomizer for PhenoDis will be ready for the application in clinics and will provide high-quality decision support for the diagnosis of rare diseases.

REFERENCES

[1]A. Arnold, J. Kern, P. Itin, M. Pigors, R. Happle, and C. Has. Acromelanosis albo-punctata: a distinct inherited dermatosis with acral spotty dyspigmentation without systemic involvement. *Dermatology*, pages 331–339, 2012.

[2]M. Baudrimont, F. Dubas, A. Joutel, E. Tournier-Lasserve, and M. Bousser. Autosomal dominant leukoencephalopathy and subcortical ischemic stroke. a clinicopathological study. *Stroke*, pages 122–125, 1993.

[3]Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B*, 57:289–300, 1995.

[4]M. Berthold, N. Cebron, F. Dill, T. Gabriel, T. Kötter, T. Meinl, P. Ohl, C. Sieb, K. Thiel, and B. Wiswedel. Knime: The konstanz information miner. *Data Analysis, Machine Learning and Applications Studies in Classification, Data Analysis, and Knowledge Organization, V*, pages 319–326, 2008.

[5]H. Bloxam, M. Day, N. Gibbs, and L. Woolf. An inborn defect in the metabolism of tyrosine in infants on a normal diet. *Biochem J*, pages 320–326, 1960.

[6]W. Bond, L. Schwartz, K. Weaver, D. Levick, M. Giuliano, and M. Graber. Differential diagnosis generators: an evaluation of currently available computer programs. *J Gen Intern Med*, 27(2):213–219, 2011.

[7]M. Borden, J. Holm, J. Leslie, L. Sweetman, W. Nyhan, L. Fleisher, H. Nadler, D. Lewis, and C. Scott. Hawkinsinuria in two families. *Am J Med Genet*, pages 52–56, 1992.

[8]S. Brass, E. Smith, J. Arboleda-Velasquez, W. Copen, and M. Frosch. Case records of the massachusetts general hospital: Case 12-2009: A 46-year-old man with migraine, aphasia, and hemiparesis and similarly affected family members. *N Engl J Med*, pages 1656–1665, 2009.

[9]B. Clayton, R. Dobbs, and A. Patrick. Leigh's subacute necrotizing encephalopathy: clinical and biochemical study, with special reference to therapy with lipoate. *Arch Dis Child*, pages 467–478, 1967.

[10]K. Drummond, A. Michael, R. Ulstrom, and R. Good. The blue diaper syndrome: familial hypercalcemia with nephrocalcinosis and indicanuria; a new familial disease, with definition of the metabolic abnormality. *Am J Med*, pages 928–948, 1964.

[11]J. Gonzãlez and M. Vãzquez Botet. Acromelanosis: A case report. *J Am Acad Dermatol*, pages 128–131, 1980.

[12]A. Hamosh, A. F. Scott, J. S. Amberger, C. A. Bocchini, and V. A. McKusick. Online mendelian inheritance in man (omim), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Research*, 33:D514–D517, 2005.

[13]Z. He, X. Luo, L. Liang, P. Li, D. Li, and M. Zhe. Merosin-deficient congenital muscular dystrophy type 1a: A case report. *Exp Ther Med*, pages 1233–1236, 2013.

[14]F. Hommes, H. Polman, and J. Reerink. Leigh's encephalomyelopathy: an inborn error of gluconeogenesis. *Arch Dis Child*, pages 423–426, 1968.

[15]J. Ip, P. Hui, M. Chau, and W. Lam. Merosin-deficient congenital muscular dystrophy (mdcmd): a case report with mri, mrs and dti findings. *J Radiol Case Rep*, pages 1–7, 2012.

[16]K. Jones, G. Morgan, H. Johnston, V. Tobias, R. Ouvrier, I. Wilkinson, and K. North. The expanding phenotype of laminin alpha2 chain (merosin) abnormalities: case series and review. *J Med Genet*, pages 649–657, 2001.

[17]S. Kamat, P. Pepmueller, and T. Moore. Triplets with systemic lupus erythematosus. *Arthritis Rheum*, 48(11):3176–80, November 2003.

[18]A. Kanwar, R. Jaswal, G. Thami, and G. Bedi. Acquired acromelanosis due to phenytoin. *Dermatology*, pages 373–374, 1997.

[19]A. Kelly, P. Lunt, F. Rodrigues, P. Berry, D. Flynn, P. McKiernan, D. Kelly, G. Mieli-Vergani, and C. TM. Classification and genetic features of neonatal haemochromatosis: a study of 27 affected pedigrees and molecular analysis of genes implicated in iron metabolism. *J Med Genet*, 38(9):599–610, November 2001.

[20]KNIME.COM AG. Konstanz information miner. `https://www.knime.org/`.

[21]S. Köhler, M. H. Schulz, P. Krawitz, S. Bauer, S. Dölken, C. E. Ott, C. Mundlos, D. Horn, S. Mundlos, and P. N. Robinson. Clinical diagnostics in human genetics and semantic similarity searches in ontologies. *The American Journal of Human Genetics*, 85:457–464, 2009.

[22]D. Leigh. Subacute necrotizing encephalomyelopathy in an infant. *J Neurol Neurosurg Psychiatry*, pages 216–221, 1951.

[23]S. Libit, R. Ulstrom, and D. Doeden. Fecal pseudomonas aeruginosa as a cause of the blue diaper syndrome. *J Pediatr*, pages 546–547, 1972.

[24]MedDRA MSSO. Medical dictionary for regulatory activities. http://www.meddra.org/.

[25]A. Niederwieser, A. Matasovic, P. Tippett, and D. Danks. A new sulfur amino acid, named hawkinsin, identified in a baby with transient tyrosinemia and her mother. *Clin Chim Acta*, pages 345–356, 1977.

[26]J.-D. Quell, M. Hastreiter, S. J. Kopetzky, J. Hoser, and F. Büttner. A new universal next generation sequencing workflow toolkit for knime. Masterpraktikum at Helmholtz Center Munich - German Research Center for Environmental Health (GmbH), Institute of Bioinformatics and Systems Biology, Neuherberg, Germany, 2012.

[27]A. Rath, A. Olry, F. Dhombres, M. M. Brandt, B. Urbero, and S. Ayme. Representation of rare diseases in health information systems: The orphanet approach to serve a wide range of end users. *Human Mutation*, 33:803–808, 2012.

[28]P. N. Robinson, S. Köhler, S. Bauer, D. Seelow, D. Horn, and S. Mundlos. The human phenotype ontology: A tool for annotating and analyzing human hereditary disease. *The American Journal of Human Genetics*, 83:610–615, 2008.

[29]A. Schiepatti, J.-I. Henter, E. Daina, and A. Aperia. Why rare diseases are an important medical and social issue. *Lancet*, 371:2039–2041, 2008.

[30]H. Siemens. Aacromelanosis albo-punctata. *Dermatologica*, pages 86–87, 1964.

[31]K. Tomoeda, H. Awata, T. Matsuura, I. Matsuda, E. Ploechl, T. Milovac, A. Boneh, C. Scott, D. Danks, and F. Endo. Mutations in the 4-hydroxyphenylpyruvic acid dioxygenase gene are responsible for tyrosinemia type iii and hawkinsinuria. *Mol Genet Metab*, pages 506–510, 2000.

[32]E. Tournier-Lasserve, M. Iba-Zizen, N. Romero, and M. Bousser. Autosomal dominant syndrome with strokelike episodes and leukoencephalopathy. *Stroke*, pages 1297–1302, 1991.

[33]UK Department of Health. Uk strategy for rare diseases. `https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/260562/UK_Strategy_for_Rare_Diseases.pdf`, 2014.

SUPPLEMENTARY MATERIAL

Orphanet terms	HPO terms	Frequency range
Occasional	Very rare	$[0, 0.25]$
	Rare	
	Occasional	
Frequent	Frequent	$]0.25, 0.75]$
	Typical	
	Variable	
	Common	
Very frequent	Hallmark	$]0.75, 1]$
	Obligate	

Table S1. Mapping of frequency terms and values to Orphanet terms.

	Whole article		Case description	Abstract	Manual
Acromelanosis (1) [11]	55 (67)		208 (214)	74 (48)	
Acromelanosis (2) [1]	25 (11)		1 (2)	41 (37)	
Acromelanosis (3) [18]	non-readable pdf		no case description	no keywords	21 (20)
Acromelanosis (4) [30]	non-readable pdf		no case description	48 (45)	22 (50)
Blue Diaper Syndrome (1) [10]	7 (6)		13 (35)	no keywords	1 (3)
Blue Diaper Syndrome (2) [23]	1496 (1521)		5155.5 (5155.5)	1575 (1568)	65 (66)
CADASIL (1) [8]	10 (69)		26 (84)	76 (71)	
CADASIL (2) [32]	1 (1)		33 (25)	25 (26)	
CADASIL (3) [2]	1 (1)		1 (1)	4 (1)	
Hawkinsinuria (1) [25]	4 (2)		128 (112)	1 (1)	
Hawkinsinuria (2) [31]	12 (12)	Case 1	19 (17)	1 (3)	
		Case 2	4 (4)		
		Case 3	18 (4)		
Hawkinsinuria (3) [7]	3 (1)	Case 1	18 (4)	29 (28)	
		Case 2	74 (18)		
Hawkinsinuria (4) [5]	703 (787)		no case description	no keywords	
Leigh Syndrome (1) [14]	190 (186)		262 (263)	93 (95)	
Leigh Syndrome (2) [9]	2 (18)	Case 1	76 (74)	12 (12)	
		Case 2	26 (24)		
		Case 3	10 (8)		
Leigh Syndrome (3) [22]	114 (113)		6 (6)	24 (25)	194 (192)
Congenital Muscular Dystrophy (1) [13]	1 (3)		14 (25)	5 (8)	
Congenital Muscular Dystrophy (2) [16]	21 (13)	Case 1	124 (160)	1 (2)	
		Case 2	409 (253)		
		Case 3	326 (241)		
		Case 4	717 (701)		
		Case 5	357 (457)		
Congenital Muscular Dystrophy (3) [15]	6 (10)		257 (287)	23 (27)	
Neonatal Hemochromatosis [19]	1 (1)	Case 1	1 (1)	96 (104)	
		Case 2	1 (1)		
		Case 3	1 (1)		
Systemic Lupus [17]	35 (14)	Case 1	2 (1)	3 (1)	
		Case 2	1 (1)		
		Case 3	2 (4)		

Table S2. Rankings of the different case studies for the four different types of symptom extraction (whole article, case description, keywords from abstract and manually extracted symptoms from the whole article). The first rank is obtained by using unweighted Phenomizer with p values. The second value in brackets is the corresponding rank for weighted Phenomizer with p values.

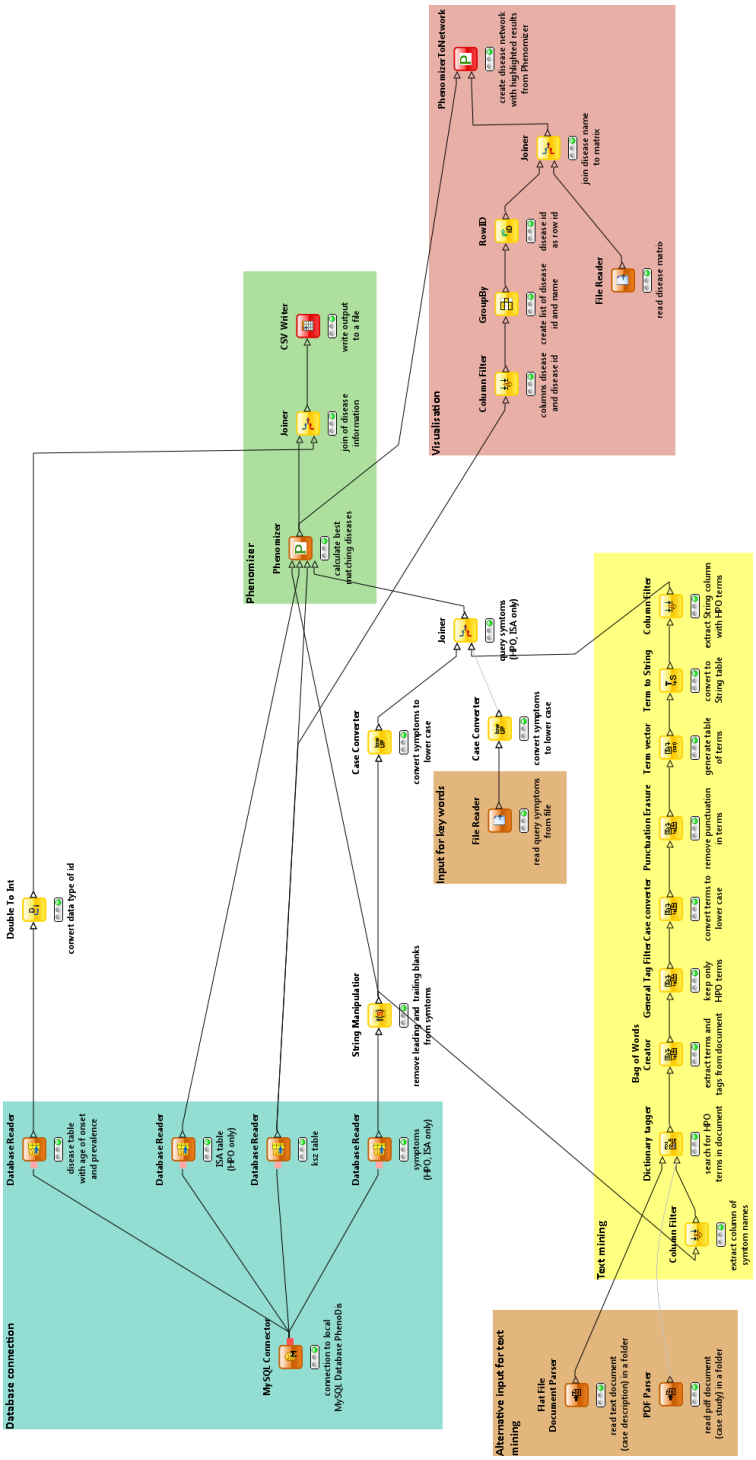

Fig. S1. Whole workflow of our project: Firstly, a connection to the local PhenoDis database is needed, also including filtering and processing of the data (blue). Secondly, there are different input possibilities to generate the query for Phenomizer (orange). In the left lower corner, two alternative inputs (represented by a grey line) for the tex tmining workflow (yellow) are shown. One can either upload a pdf document (e.g. the whole article) or a text document (e.g. the case description). In the subsequent text mining workflow, the document is searched for symptom names and synonyms of PhenoDis. Alternatively, a file containing symptom names for a query can be loaded ("Input for keywords"). These nodes can be directly connected to the Joiner (shown by a grey line). The actual Phenomizer algorithm uses tables from the database and a query as input and computes the best matching diseases (green). Afterwards, additional information about the diseases (prevalence, age of onset) are added to the output of Phenomizer. Finally, the results are written to a file. Furthermore, a visualisation of the results is possible (red). Here, a network of diseases is generated based on a matrix of similarity scores between all pairs of diseases. In this network, the top scoring results for a query are highlighted.

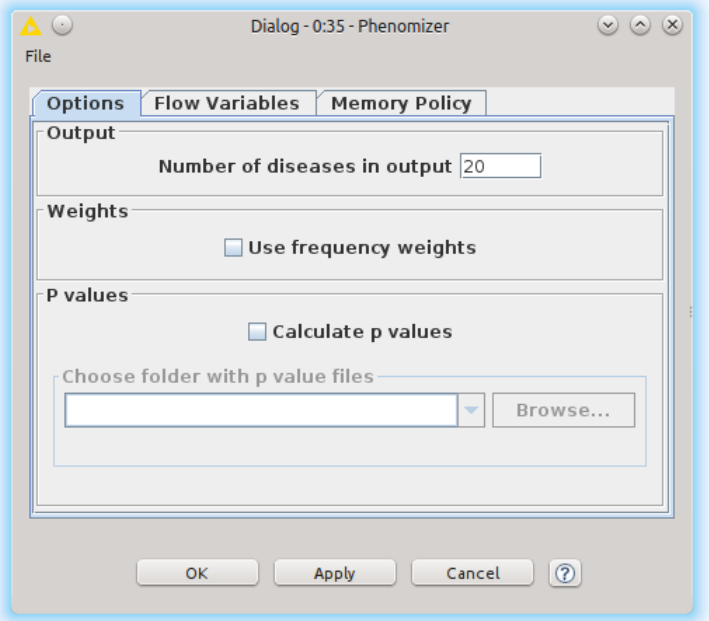

Fig. S2. Screenshot of the node dialog of the Phenomizer node in KNIME. The node provides several input options: adjusting of the number of diseases in the output, option to use frequency weights (one-sided), option to use p values (path to folder with files of precalculated score distributions needed).

Most Similar Diseases - 0:15 - Phenomizer

File

Table "default" - Rows: 20 | Spec - Columns: 5 | Properties | Flow Variables

Row ID	disease_id	disease	score	p_value	significance
Row 1	5644	Neonatal hemochromatosis	3.17	0	***
Row 2	4805	Dehydrated hereditary stomatocytosis	2.597	0	***
Row 3	2205	Atypical hemolytic uremic syndrome with thrombomodulin anomaly	2.557	0	***
Row 4	7390	Atypical hemolytic uremic syndrome with C3 anomaly	2.557	0	***
Row 5	7391	Atypical hemolytic uremic syndrome with MCP/CD46 anomaly	2.557	0	***
Row 6	7392	Atypical hemolytic uremic syndrome with B factor anomaly	2.557	0	***
Row 7	7394	Atypical hemolytic uremic syndrome with I factor anomaly	2.557	0	***
Row 8	6512	Congenital bile acid synthesis defect type 3	2.477	0	***
Row 9	501	Budd-Chiari syndrome	2.341	0	***
Row 10	9827	BILE ACID SYNTHESIS DEFECT, CONGENITAL, 4	2.258	0	***
Row 11	166	Familial isolated congenital asplenia	2.231	0	***
Row 12	5312	Atypical hemolytic uremic syndrome with DGKE deficiency	2.058	0	***
Row 13	5213	Propionic acidemia	2.052	0	***
Row 14	6498	Erythropoietic protoporphyria	2.051	0	***
Row 15	4093	Fetal parvovirus syndrome	2.042	0	***
Row 16	6347	Hereditary hemorrhagic telangiectasia	2.016	0	***
Row 17	3991	Isolated polycystic liver disease	2.005	0	***
Row 18	8521	Klatskin tumor	1.983	0	***
Row 19	2541	Beta-thalassemia - X-linked thrombocytopenia	1.978	0	***
Row 20	7224	Renal cysts and diabetes syndrome	1.954	0	***

Fig. S3. Screenshot of the output of Phenomizer. The 20 top scoring diseases for the symptoms coming from the article about Neonatal Hemochromatosis [19] are listed. The output has five columns: disease_id refers to the PhenoDis id, disease refers to the name from PhenoDis, score is the calculated similarity score between the query and the disease, p value gives the corresponding p value for this similarity score, significance indicates whether the p value is smaller than a certain significance level or not. The results are sorted in ascending order according to the p values and in decreasing order according to the similarity scores.

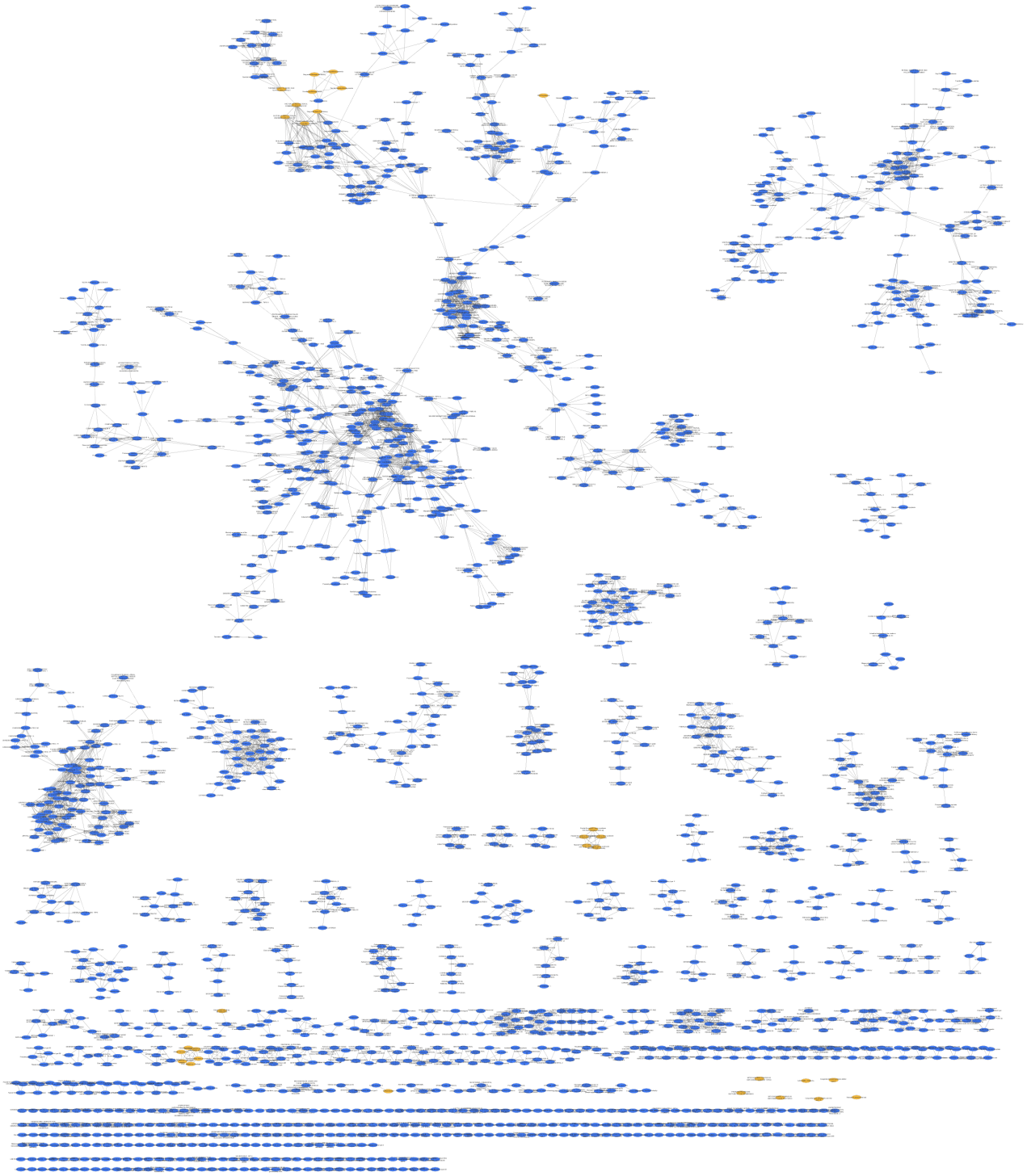

Fig. S4. Network of the diseases from the PhenoDis database. Each node represents a disease. Edges are drawn between pairs of diseases obtaining a similarity score > 4.0. Yellow-coloured nodes represent the top scoring diseases from the output of the unweighted Phenomizer without p values. The input for the query are symptoms from the article about Neonatal Hemochromatosis [19].

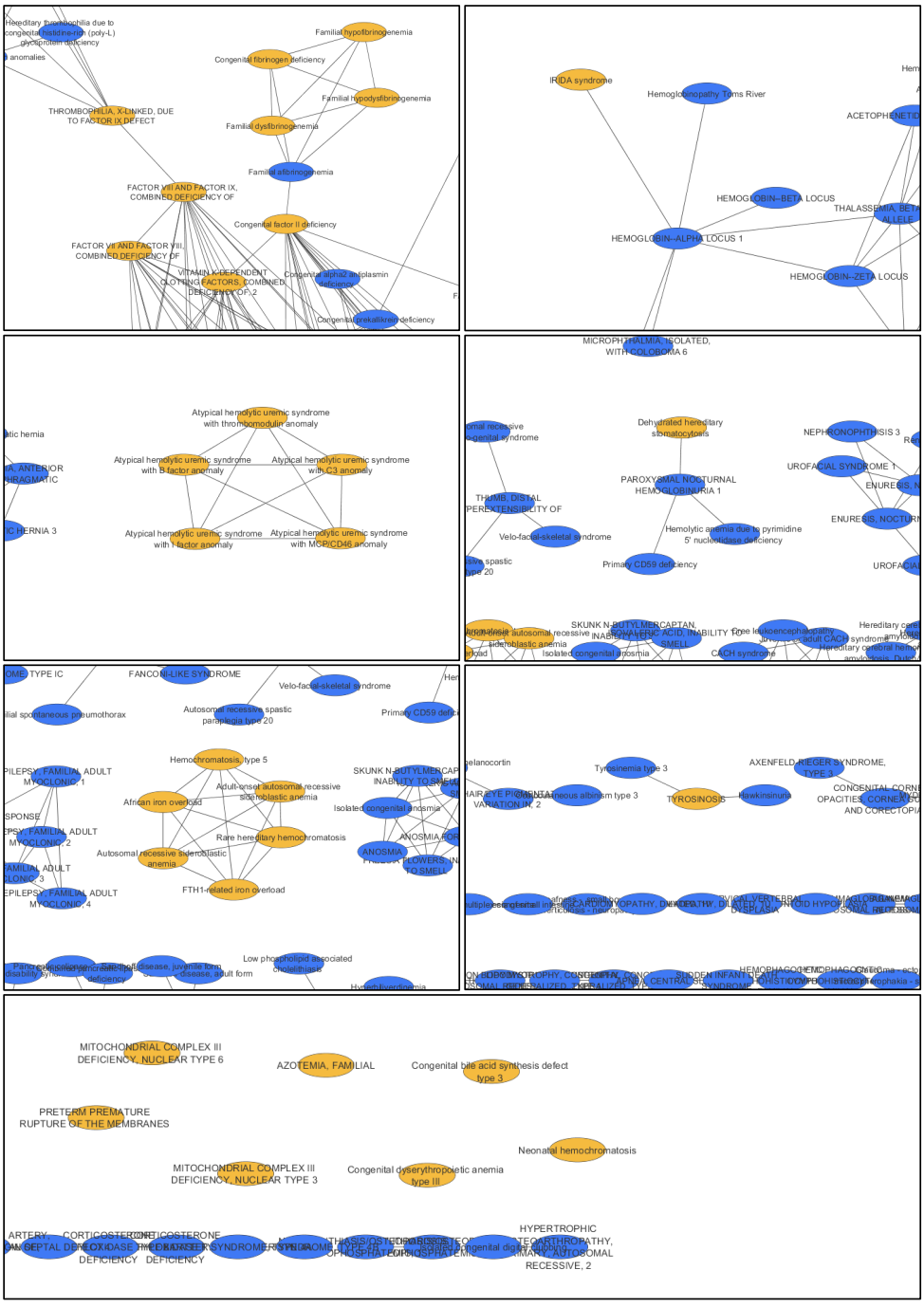

Fig. S5. Parts from the network in fig S4. All 30 top scoring diseases of the output from Phenomizer for the article about Neonatal Hemochromatosis [19] are shown.

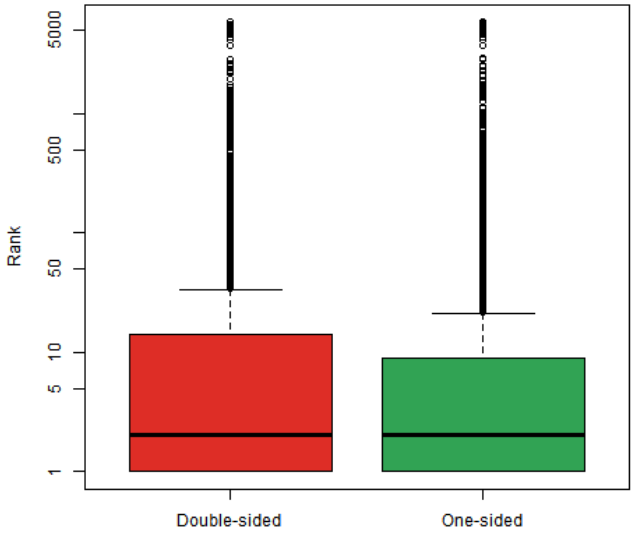

Fig. S6. Rankings of the OMIM diseases for the two different implementations of the weighted Phenomizer without p values: Calculation of the double-sided and the one-sided weighted similarity score. Comparing the results shows that the weighted Phenomizer accomplishes a higher accuracy for the one-sided weighting than for the double-sided one.

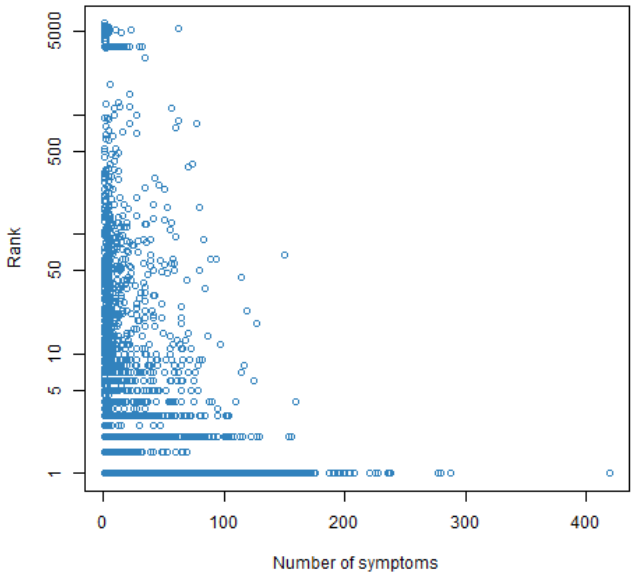

Fig. S7. Correlation between the number of annotated symptoms in PhenoDis and of the ranks. The ranks were obtained by the unweighted Phenomizer with p values for the diseases of the OMIM data. There is a significant negative correlation between the number of symptoms and the rank: the correlation coefficient is -0.17 with p value 0.

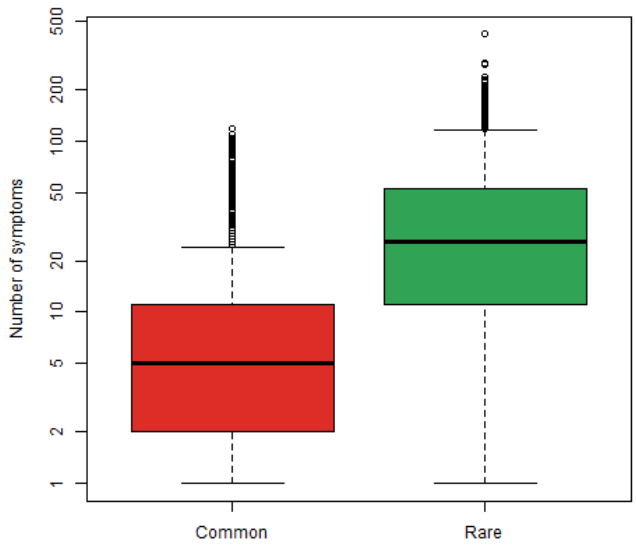

Fig. S8. Number of symptoms annotated to common and rare diseases in PhenoDis. Comparing the results for PhenoDis shows that rare diseases are in general annotated with more symptoms than common diseases.

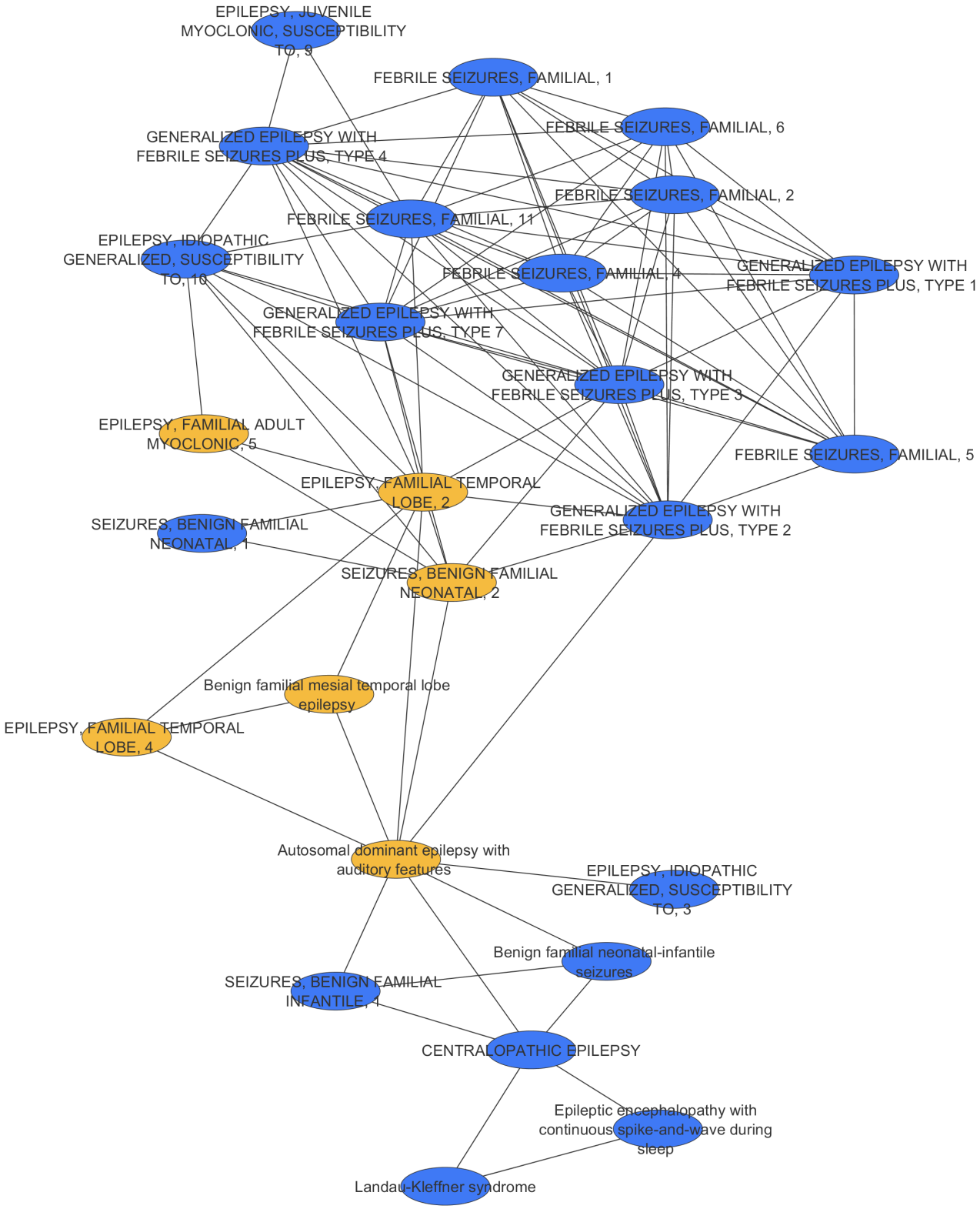

Fig. S9. Part of a network created by the visualisation node PhenomizerToNetwork. The symptoms from the whole article about CADASIL [32] are used as input for the unweighted Phenomizer without p values. The yellow nodes are part of the 30 top scoring diseases for this query. The connected component shown here consists of related diseases of different kinds of epileptic seizures. Six of the 26 nodes are listed in the output of Phenomizer. But because of the relatedness of this family of diseases, all diseases are a plausible diagnosis matching the symptoms of the query.