Towards the Automated Identification of Orphan Diseases From Case Descriptions

Christian Rohrdantz*, Andreas Stoffel*, Franz Wanner*, Martin Drees†
*Department of Computer & Information Science, University of Konstanz, Germany
Email: firstname.lastname@uni-konstanz.de
†coliquio GmbH, Konstanz, Germany

***Abstract*—Orphan diseases are very rare diseases that are not well-known to many medical doctors. Patients suffering from them often remain without the correct diagnosis. Yet, there is a potential that advice-seeking doctors, posting medical case descriptions in web forums, may be automatically given a hint to matching orphan diseases. In this work-in-progress paper, we investigate opportunities and issues for an automated identification of orphan diseases in medical case descriptions through text mining and data analytics.**

Keywords–Data Analytics; Text Mining; Medical Data Analytics.

I. INTRODUCTION

It is the daily work of most medical doctors to examine patients and then combine observations on clinical signs and symptoms with their knowledge and experience in order to arrive at diagnoses. Accurate and timely diagnoses are crucial for initiating successful treatments. Yet, there are cases where medical doctors are confronted with patients showing symptoms that do not fit well into the known patterns. In these cases, physicians consult literature and seek the advice of experienced colleagues.

But what if the disease of the patient is just so extremely rare that only a handful of experts worldwide would be able to identify it based on the given clinical signs? There is a quite high chance that these cases end up with unspecific or wrong diagnoses and do not get the optimal treatment. Yet, it is known that there are quite a number of such rare, so-called orphan diseases.

In recent years, doctors have increasingly made use of medical web forums in order to seek advice from colleagues and discuss cases. In Germany, the largest and most active online community of medical doctors is coliquio [1]. It is experiencing a fast growth and has currently already more than 125,000 members. Without a doubt, it would be of great value if the case descriptions of advice-seeking physicians could be automatically matched with the known orphan diseases. These physicians could then be hinted by the system to the corresponding orphan disease in cases where a match seems likely.

Consequently, in this paper we describe work-in-progress towards such an automated identification of orphan diseases. The contribution of this initial study is twofold. First, we investigate how freely available, stuctured resources of medical knowledge, like orphanet [2], [3], and large repositories of medical texts, like the Wikipedia Portal Medicine [4] , can be exploited for our purpose. Second, we present a statistical method for the extraction of contextual knowledge and terminology, and show that it yields a lot of relevant information that could not be gathered from common dictionaries and medical ontologies.

The remainder of this paper is organized as follows. First, in Section II, we provide the necessary background information on rare or orphan diseases and the related work in this area, before we describe the data sources used in Section III. Next, in Section IV, we introduce our novel approach of leveraging the described data for the potential detection of orphan diseases from medical case descriptions. In Section V, we present first results, evaluate the performance, and identify issues and opportunities. Finally, we draw conclusions and identify key challenges setting the agenda for future work in Section VI.

II. RELATED WORK

Orphan or rare diseases fit into the broader context of research regarding rare events [5] and events in text data [6], but to date have not been treated in these areas.

The definitions of orphan disease vary slightly in the literature. The European Organisation for Rare Diseases (EURORDIS) states on their Websites:

"A rare disease, also referred to as an orphan disease, is any disease that affects a small percentage of the population. Most rare diseases are genetic, and are present throughout a person's entire life, even if symptoms do not immediately appear. In Europe, a disease or disorder is defined as rare when it affects less than 1 in 2000 citizens." [7].

Despite of the rareness of individual diseases, the overall quantity of affected people is still quite high:

"There are more than 6000 rare diseases. On the whole, rare diseases may affect 30 million European Union citizens." [8]

The coverage of orphan diseases in standard terminologies is very limited [9]. Rath et al. [10] state that only 446 orphan diseases have a specific code in the ICD10 disease classification, which most European countries use in their health information systems.

In general, text mining for clinical medical records is an important field of research [11], but there is few work on symptom or disease identification. Koeling et al. [12] manually annotate symptoms in patient records and provide statistical information on the frequency distribution of symptoms. They come to the conclusion that "there is great variation in the expressions used to describe the same symptom". Data from orphanet has been used exploiting the given mapping between diseases and disease-causing genes [13], but not for text mining purposes. Our approach fills a clear gap in the current research.

III. DATA

A. Orphadata

Orphanet provides structured textual data on orphan diseases and indicative clinical signs as part of their orphadata

 ISBN: 978-1-61208-423-7

service [14]. We made use of the XML version of the data in German. The data contains information about 2689 different orphan diseases and their clinical signs. Overall, the data contains 1362 different clinical signs and information on their frequency for different diseases. Moreover, the clinical signs are organized hierarchically in a thesaurus structure from rather general to more specific signs. Each clinical sign is typically described by one or more synonyms or alternative expressions. For example, one of the clinical signs is named "Nausea/vomiting/regurgitation/merycism/hyperemesis". We will refer to each of these alternative expressions as *symptoms*. For each orphan disease, different clinical signs may have three different frequency values: *very frequent*, *frequent*, and *occasional*. While the data is available for different languages, in this initial study we use the German version only.

B. Wikipedia Portal Medicine

The Wikipedia Portal Medicine constitutes a rich body of diverse textual medical information. We leverage the German version of this resource in order to automatically extract context knowledge and feed it into our text mining models.

IV. Mining and Modeling Medical Knowledge from Text

One of the services orphanet provides is that a user can select different clinical signs from the controlled thesaurus through a web interface and retrieve potentially matching orphan diseases. The big challenge we face, however, is to automatically identify mentions of clinical signs in medical case descriptions. In only very few cases, physicians use explicitly and exactly the terminology given in controlled vocabularies like the orphanet thesaurus. Mostly, they will use either inflected word forms, alternative wordings, varying multi-word expressions, paraphrasing or abbreviations. Our approach consists in learning the alternative terminology applying advanced statistical methods to large text repositories, such as the Wikipedia Portal Medicine. The advantage is that such a source contains expressions as they are actually used by phycisians rather than controlled idealized language use. For the mining of medical knowledge from texts, we proceed different consecutive steps.

A. Step 1: Identifying Descriptive Contexts

As mentioned before, each clinical sign in the orphadata is described by a set of symptoms. In order to get hold of textual contexts describing symptoms, we query Wikipedia. For each symptom, our first attempt is finding an article where the title exactly matches the given symptom. Such an article has basically been written to describe the symptom and consequently we consider all of the text in the article to be related to the symptom. For us, it constitutes what we define as a *descriptive context*. If, however, there is no matching article, we make use of the common search capability of Wikipedia. We use the symptom as a query and then sift through the retrieved articles. For each of these articles, we first check whether it belongs to the category "medicine" or one of its more than 400 subcategories. From each article meeting this criterion, we extract those paragraphs, where the symptom appears and save them as descriptive contexts.

B. Step 2: Extracting Knowledge from Descriptive Contexts

Next, from the available descriptive contexts we build two kinds of data co-occurrence tables. First, for each noun we count how often it occurs within the descriptive contexts of each symptom. This gives us a noun-symptom co-occurrence table. Next, for each pair of nouns, we count how frequently they co-occur within descriptive contexts. This gives us a noun-noun co-occurrence table for symptom contexts. From now on, we will refer to nouns within the tables as *descriptors*.

C. Statistics-based Identification of Relevant Descriptors

Next, we perform a statistical analysis of the co-occurrence tables. First, we aggregate the descriptor counts for all symptoms belonging to the same clinical sign. Next, we extract those descriptors that are highly correlated with individual clinical signs. The assumption is that if these descriptors can be identified in a medical case description, they will be likely to point to the correlated clinical sign.

V. Preliminary Results & Evaluation

In order to gain a better feeling for the feasibility of an automatic detection, we systematically analyze both the information contained in the orphadata and that extracted from Wikipedia. We perform different statistical analyses in order to learn more about potential issues and opportunities.

A. Knowledge Extraction from Orphadata

The listing of orphan diseases and their clinical signs builds the backbone of our approach. The nature of this data may therefore impose limitations on the overall proceeding. In a first step, we have to examine and evaluate this resource.

An orphan disease contained in orphadata has between as few as one and as many as 180 different clinical signs. On average an orphan disease contains 19.5 clinical signs, with a standard deviation of 15.7. Yet, the distribution is somewhat skewed and the peak is with 9 clinical signs per disease, see Figure 1. At the lower end the data sparsity may be an issue for the identification of orphan diseases: 32 orphan diseases have only one clinical sign each, and 58 have only two signs each. It is questionable whether the automated detection is feasible for these cases as it will be based on a lightweight evidence. Orphan diseases at the other side of the range, the ones having a plethora of clinical signs, are challenging for the analysis, too. Yet, there is a chance to narrow the list of clinical signs down to the most indicative ones. The disease with most clinical signs has 180 of them, out of which 48 are classified as *very frequent*, 50 as *frequent*, and 82 as *occasional* only. It can be observed as a general tendency that diseases with more signs tend to have a disproportionately high amount of *occasional* signs.

A clinical sign contained in orphadata points to as few as one and as many as 987 different orphan diseases. On average a clinical signs points to 41.2 different orphan diseases, with a standard deviation of 71.5. Again, we are confronted with a quite skewed distribution with a peak at two different diseases per sign, see Figure 2. At the lower end of the scale, we find clinical signs that are quite specific and indicative for certain orphan diseases. For example, 66 signs point to exactly one disease each, and 73 signs point two different diseases each. The discriminatory power of these clinical signs within the set of orphan diseases can be considered to be quite high.

 ISBN: 978-1-61208-423-7

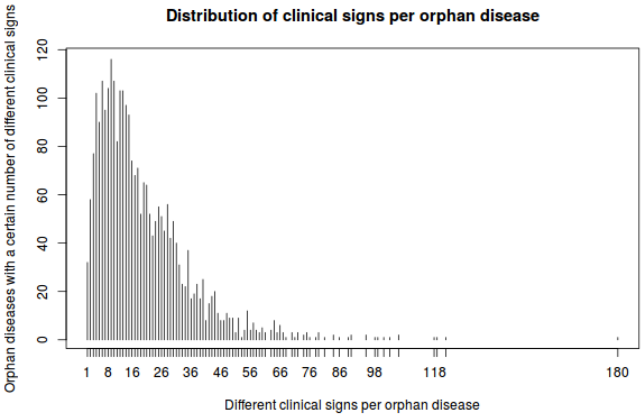

Figure 1. Skewed distribution: some orphan diseases have far more different clinical signs than others.

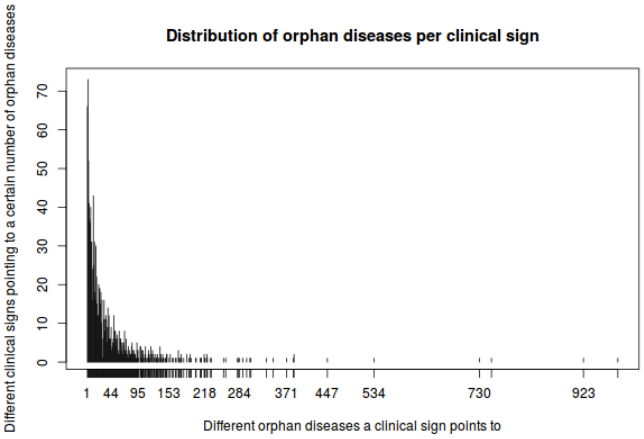

Figure 2. Skewed distribution: some clinical signs point to far more different orphan diseases than others.

At the other end of the range, we find widely spread signs like the one named "Intellectual deficit/mental/psychomotor retardation/learning disability", which points to 987 different diseases. When omitting orphan diseases for which clinical signs are *occasional* only, the tendency remains the same. The most widely spread sign occurs for 876 different diseases *frequently* or *very frequently*. Still, those clinical signs pointing to a wide range of different diseases may be quite useful for the higher-level classification whether a patient might suffer from an orphan disease or not. For a distinction within the set of orphan diseases, more specific, hardly spread signs are useful.

With $n = 2689$ different orphan diseases we can make $(n * (n-1))/2 = 3,614,016$ pairwise comparisons. In particular, we can determine for each pair of diseases in how many clinical signs they coincide and in how many they are distinct. Figure 3 provides a visual summary of performing all pairwise comparisons. The strong left shift of the resulting distribution clearly shows that for almost all of these pairwise comparisons, the corresponding orphan diseases are quite distinct: they share only very few clinical signs (x-axis) while there are many clinical signs in which they can be distinguished (y-axis). This shows that most orphan diseases have quite unique combinations of signs, which makes an automated distinction become very realistic.

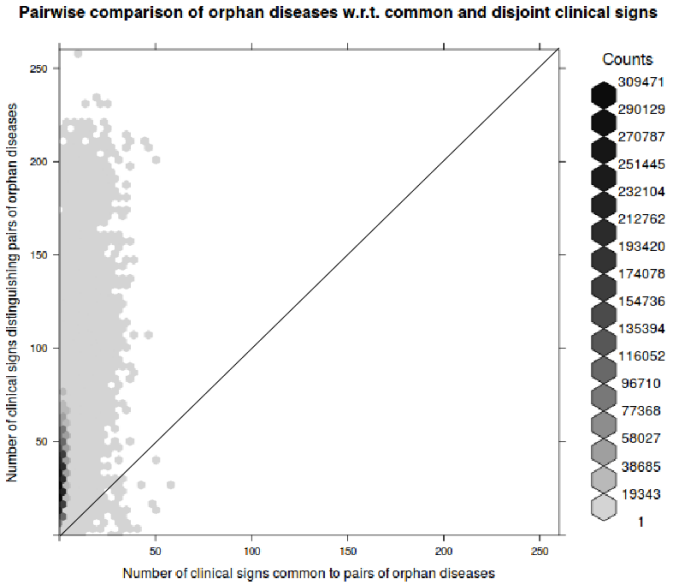

Figure 3. The lion's share of the data is located to the upper left of the diagonal, i.e., almost all pairs of diseases differ in more clinical signs than they have in common.

B. Knowledge Extraction from Wikipedia

The 1362 clinical signs from orphadata contain more than 2500 symptoms. Currently, for 37.2% of the symptoms Wikipedia articles exist, for 15,8% of the symptoms related articles can be retrieved, and for the remaining 47% no articles are found. Overall, information from 2479 Wikipedia pages was gathered. Yet, as described, for a considerable number of symptoms no information was available. This implies that for 423 clinical signs, roughly one third of all signs, we do not dispose of descriptive contexts.

From the available contexts, descriptors were extracted as described in Section IV. When investigating the extracted descriptors, it becomes evident that they are typically very useful and can be divided into 8 categories. For a better illustration, Table I provides the top 50 correlations to clinical signs together with the categories they fall in. The categories and their relative frequencies within the top 50 descriptors:

a) SYNM (28%): Synonyms, e.g., *Lichtscheu* for *Photophobie*. As in this case, often one synonym has a German origin, while the other is of Latin or Greek provenance.

b) ORTH (6%): Orthographical variations, e.g., *Hydrocephalus* for *Hydrozephalus*.

c) ABBR (4%): Abbreviations, e.g., *AVSD* for *atrioventrikulärer Kanal*.

d) DISE (8%): Diseases that relate to the given symptom, e.g., *KBG-Syndrom* for *EEG-Anomalien*.

e) THRP (2%): Terms indicating a common therapy for the given clinical sign.

 ISBN: 978-1-61208-423-7

f) GNRL (10%): Hypernyms or otherwise more general terms with and without morphological relation, e.g., *Volvulus* for *Magenvolvulus*, or *Gliedergürteldystrophien* for *Klauenzehe/Beugekontrakturen der Zehen*.

g) RELA (34%): Clearly related terms that do not fall into any of the other previous categories, e.g., *Fusionsgene* for *interstitielle Deletion/subtelomere Mikrodeletion*.

h) ERRD (8%): Errors typically due to wrong data. In all of the investigated cases, we could trace them back to the erroneous retrieval of an unrelated article in Wikipedia.

It can be concluded that a number of descriptors could potentially have been discovered by other means as well. For example, spell-checkers could have uncovered orthographical variations. Synonyms and abbreviations could potentially have been retrieved from digital dictionaries. Specialized ontologies could have helped in extracting related diseases and therapies. That makes a total of about 48% of descriptors, for which there is hope to obtain them by alternative means.

For the 10% of more general terms (GNRL), however, it is doubtful whether these could have been gathered from controlled vocabularies. For the 34% of remaining related terms (RELA), finally, there is basically no other way than learning them from data. Our approach does a very good job with respect to this and on top it extracts descriptors from all of the other categories with the same proceeding. The automated method yields whole semantic fields with a precision of more than 90% and provides more than a third more descriptors than available by the most optimistic usage scenario of traditional resources. It is hard to estimate a reasonable recall value, though. Finally, apart from a word lemmatization processing step, our method is language-agnostic and can readily be transferred to other natural languages.

VI. Conclusion & Future Work

This work-in-progress paper lays the foundation for the automated extraction of orphan diseases from medical case descriptions and uncovers a number of challenges and limitations mainly regarding the available data.

For some orphan diseases, orphanet lists only one or two symptoms. In these cases, we have a limited and uncertain foundation for identification. In addition, for one third of all clinical signs we could not retrieve descriptive contexts from Wikipedia. This limits the identification of these signs in medical case descriptions to mere exact matches.

While in some cases, due to nature of the data or the lack of available data, the automated identification of orphan diseases is currently quite challenging, for the majority of orphan diseases we indeed do see a very good chance. Moreover, the used data sources are permanently growing and being improved, which will ease the identification and put it in on more solid ground in the future. Regarding the extraction of descriptors, the first results can be considered as very promising. The precision is above 90%. In the future, we plan to extract more than just nouns as descriptors, namely words with other parts-of-speech, word ngrams and phrases. This will increase the recall further. Finally, we will start experimenting with different ways of incorporating the descriptors for an automated identification of orphan diseases within case descriptions posted to the medical community coliquio.

TABLE I. THE 50 STRONGEST CORRELATIONS BETWEEN CLINICAL SIGNS AND DESCRIPTOR WORDS TOGETHER WITH THE CATEGORY THEY FALL IN.

Clinical Sign (with orphanet id)	Descriptor	Cat.
47800 Kryoglobulinämie	Kryoglobuline	RELA
20360 Trommelschlegelfinger	Trommelschlägelfinger	ORTH
5720 Photophobie	Lichtscheu	SYNM
46560 Kniescheibenverrenkung	Patellaluxation	SYNM
23020 Hypohidrose/...	Anhidrose	SYNM
7150 Blepharophimose/...	Blepharophimose-Syndrom	RELA
21320 Anom. d. unt. Extremitäten/...	Epiphyseodese	THRP
15640 Pectus carinatum	Kielbrust	SYNM
33350 Ausweitung der Bronchien/...	Bronchiektasen	RELA
52480 interstitielle Deletion/...	Fusionsgene	RELA
52540 Chromosomenbrüchigkeit	Nijmegen-Breakage-Syndrom	DISE
41870 Galaktorrhö	Milchfluss	SYNM
43140 EEG-Anomalien	KBG-Syndrom	DISE
3700 Kinngrübchen/Kinnspalte	Grübchen	GNRL
22320 Klauenzehe/...	Gliedergürteldystrophien	GNRL
41750 vorzeitige Pubertät	Pubertas	SYNM
15400 überzählige Mamillen/...	Milchleiste	SYNM
4260 Melanose der Iris/...	Melanosis	ORTH
49680 Vitamin B3/PP-Mangel	Nicotinsäure	SYNM
35270 .../Raynaud-Phänomen/...	Raynaud-Syndrom	SYNM
23330 negatives Nikolski-Zeichen	Pemphigus	DISE
17880 persistierender Urachus/...	Allantois	RELA
23060 Hautdehnungsstreifen/Striae	Dehnungsstreifen	GNRL
27630 Darmverschluss/...	Ileus	SYNM
43200 Gangstörung/auffälliger Gang	Gangbild	RELA
10490 Ankyloglossie/...	Zungenbändchen	RELA
42450 Hydrozephalus	Hydrocephalus	ORTH
35480 Ödem der Beine/...	Frakturen	ERRD
49260 Hyperkalziurie	Nephrokalzinose	RELA
5060 Glaskörpertrübungen/...	Vitrektomie	RELA
23190 chron. Infektion der Haut/...	Ulcus	RELA
54210 Durst	Durstgefühl	RELA
34500 atrioventrikulärer Kanal	AVSD	ABBR
18880 kutanes/amniotisches Band/...	Amniotisches-Band-Syndrom	DISE
12250 überzählige Zähne/Polydontie	Hyperdontie	SYNM
49140 Hypokaliämie	Kalium	RELA
26420 Magenvolvulus	Volvulus	GNRL
2600 Kopfhaut/Schädeldefekt	Skalp	SYNM
2600 Kopfhaut/Schädeldefekt	Kopfschwarte	SYNM
2600 Kopfhaut/Schädeldefekt	Skalpieren	ERRD
41150 Kropf	Kropfmilch	ERRD
44450 Anomale Muskel-Biopsie/Muskelenzyme/CPK/LDH/...	Enzym	GNRL
same as above	Blutentnahme	RELA
same as above	Röhrchen	ERRD
21680 Knickfuß	Knöchel	RELA
21280 tarsale Anomalie/Fusion/...	Verschmelzung	SYNM
23500 Pigmentanomalien der Haut	FA-Patienten	RELA
50900 vaskulärer Tumor	EHE	ABBR
44500 Fasziitis	Faszien	RELA
24200 Lanugo/Wollhaar	Lanugobehaarung	RELA

Acknowledgment

This work has partly been funded by the research project "Visual Analytics of Text Data in Business Applications" at the University of Konstanz. We would like to thank the Vidatics GmbH for providing the software implementations.

References

[1] "coliquio," Available on: https://www.coliquio.de/ [Date accessed: 2015-05-26].

[2] "Orphanet: The Portal for Rare Diseases and Orphan Drugs." Available on: http://www.orpha.net/consor/cgi-bin/index.php [Date accessed: 2015-02-27].

[3] 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*, vol. 33, no. 5, pp. 803–808, 2012.

 ISBN: 978-1-61208-423-7

[4] "Wikipedia Portal Medicine (German Version)." Available on: https://de.wikipedia.org/wiki/Portal%3AMedizin [Date accessed: 2015-02-04].

[5] S. Uryasev and P. M. Pardalos, *Stochastic optimization*. Springer Science & Business Media, 2001, vol. 54.

[6] F. Wanner, A. Stoffel, D. Jäckle, B. C. Kwon, A. Weiler, and D. A. Keim, "State-of-the-Art Report of Visual Analysis for Event Detection in Text Data Streams," in *EuroVis - STARs*, R. Borgo, R. Maciejewski, and I. Viola, Eds. Swansea, UK: Eurographics Association, 2014, pp. 125–139.

[7] "EURORDIS, Rare Diseases Europe: What is a rare disease?" Available on: http://www.eurordis.org/content/what-rare-diseas [Date accessed: 2015-03-03].

[8] "EURORDIS, Rare Diseases Europe: About Rare Diseases," Available on: http://www.eurordis.org/about-rare-diseases [Date accessed: 2015-03-03].

[9] K. W. Fung, R. Richesson, and O. Bodenreider, "Coverage of rare disease names in standard terminologies and implications for patients, providers, and research," in *AMIA Annu Symp Proc*, vol. 564, 2014, p. 570.

[10] A. Rath, B. Bellet, A. Olry, C. Gonthier, and S. Aymé, "How to code rare diseases with international terminologies?" *Orphanet Journal of Rare Diseases*, vol. 9, no. Suppl 1, p. O11, 2014.

[11] X. Zhou, H. Han, I. Chankai, A. Prestrud, and A. Brooks, "Approaches to text mining for clinical medical records," in *Proceedings of the 2006 ACM symposium on Applied computing*. ACM, 2006, pp. 235–239.

[12] R. Koeling, J. Carroll, A. R. Tate, and A. Nicholson, "Annotating a corpus of clinical text records for learning to recognize symptoms automatically," in *Proceedings of the 3rd Louhi Workshop on Text and Data Mining of Health Documents*, 2011, pp. 43–50.

[13] S. Köhler *et al.*, "Construction and accessibility of a cross-species phenotype ontology along with gene annotations for biomedical research." *F1000Research*, vol. 2, Feb. 2013. [Online]. Available: http://dx.doi.org/10.12688/f1000research.2-30.v1

[14] "Orphadata: Free access data from Orphanet. © INSERM 1997." Available on: http://www.orphadata.org/ [Date accessed: 2015-02-04].

 ISBN: 978-1-61208-423-7