Phenotypic similarity for rare disease: ciliopathy diagnoses and subtyping

Xiaoyi Chen, Nicolas Garcelon, Antoine Neuraz, Katy Billot, Marc Lelarge, Thomas Bonald, Hugo Garcia, Yoann Martin, Vincent Benoit, Marc Vincent, et al.

Phenotypic similarity for rare disease: ciliopathy diagnoses and subtyping

Xiaoyi Chen[a], Nicolas Garcelon[b], Antoine Neuraz[a, c], Katy Billot[d, h], Marc Lelarge[e], Thomas Bonald[f], Hugo Garcia[d, h], Yoann Martin[d, h], Vincent Benoit[b], Marc Vincent[b], Hassan Faour[b], Maxime Douillet[b], Stanislas Lyonnet[g, h, i], Sophie Saunier[d, h], Anita Burgun[a, c, h]

[a] INSERM UMR1138, Centre de Recherche des Cordeliers, Team 22, Paris, France.
[b] Institut Imagine, Paris Descartes University-Sorbonne Paris Cité, Paris, France
[c] Department of Medical Informatics, Necker-Enfants Malades Hospital, Assistance Publique - Hôpitaux de Paris (AP-HP), Paris, France
[d] INSERM UMR1163, Institut Imagine, Laboratory of Inherited Kidney Diseases, Paris, France
[e] INRIA-ENS, Paris, France
[f] Telecom ParisTech, Paris, France
[g] INSERM UMR1163, Institut Imagine, Laboratory of Embryology and Genetics of Congenital Malformations, Paris, France.
[h] Paris Descartes University Sorbonne Paris Cité, Paris, France.
[i] Department of genetics, Necker-Enfants Malades Hospital, Assistance Publique - Hôpitaux de Paris (AP-HP), Paris, France.

Keywords: deep phenotyping, patient similarity; phenotypic similarity; rare disease; ciliopathies

Highlights

- In the context of complex rare disease, both diagnoses and subtyping are challenging tasks.
- Patient similarity with deep phenotyping from various data sources, including EHRs and research data, can help precise diagnosis and better subtyping.
- Combination of multiple modalities of clinical data (narratives and quantitative data) can achieve better performance of phenotyping.

Abstract

Rare diseases are often hard and long to be diagnosed precisely, and most of them lack approved treatment. For some complex rare diseases, precision medicine approach is further required to stratify patients into homogeneous subgroups based on the clinical, biological or molecular features. In such situation, deep phenotyping of these patients and comparing their profiles based on subjacent similarities are thus essential to help fast and precise diagnoses and better understanding of pathophysiological processes in order to develop therapeutic solutions. In this article, we developed a new pipeline of using deep phenotyping

to define patient similarity and applied it to ciliopathies, a group of rare and severe diseases caused by ciliary dysfunction. As a French national reference center for rare and undiagnosed diseases, the Necker-Enfants Malades Hospital (Necker Children's Hospital) hosts the Imagine Institute, a research institute focusing on genetic diseases. The clinical data warehouse contains on one hand EHR data, and on the other hand, clinical research data. The similarity metrics were computed on both data sources, and were evaluated with two tasks: diagnoses with EHRs and subtyping with ciliopathy specific research data. We obtained a precision of 0.767 in the top 30 most similar patients with diagnosed ciliopathies. Subtyping ciliopathy patients with phenotypic similarity showed concordances with expert knowledge. Similarity metrics applied to rare disease offer new perspectives in a translational context that may help to recruit patients for research, reduce the length of the diagnostic journey, and better understand the mechanisms of the disease.

1. Introduction

There are about 7000 types of rare diseases. Although individual rare diseases by definition affect few people, e.g. in Europe, a disease is defined as rare when it affects less than 1 in 2000 citizens, and in the United Stated, it is defined as rare when it affects fewer than 200,000 American people, cumulatively rare diseases have a major impact on public health (Franco, 2013; Dharssi et al., 2017). Meanwhile, the majority of the clinicians lack knowledge of these diseases, resulting in delayed diagnosis for many patients (Blöß et al., 2017; Global rare disease commission, 2018) (SHIRE, 2016). In such situation, a promising solution consists in precise and comprehensive phenotyping for the patient suspected to suffering from a rare and undiagnosed disease (Robinson, 2012), and comparing the patient's profile with similar cases recorded in the clinical data warehouses developed in the specialized rare disease centers (Garcelon et al., 2018). Patient similarity with deep phenotyping is therefore essential to help fast and precise diagnoses of rare diseases.

Patient similarity is also crucial to achieve precision medicine and stratify patients into clinically meaningful subgroups (Parimbelli et al., 2018). The similarity can be defined based on patients' molecular (genomics, transcriptomics or metabolomics) or clinical characteristics. For some complex rare diseases, precision medicine approach is required as well for subtyping objective. Our particular interest lies on such a group of rare and severe diseases, called ciliopathies, which are caused by ciliary dysfunction. As cilia are important in guiding the process of development, their dysfunction can lead to diseases with a large spectrum of clinical features ranging from embryofetal lethality, through "classic" individual organ malformation to multisystemic defects (Powles-Glover, 2014). More precisely, ciliopathies can affect nearly all organs, mainly kidneys, eyes, brain, bones and liver; and the associations of these clinical features define about 30 rare syndromes, including

nephronophthisis, Joubert syndrome, Bardet-Biedl syndrome etc., affecting in total about one per 2000 people. Over 200 ciliopathy-associated genes have been determined as mutated in ciliopathy patients. Genetic analyses of ciliopathies revealed a vast clinical variability and a broad genetic heterogeneity as: (i) mutations of the same disease-causing gene can result in distinct clinical entities and, conversely, (ii) mutations in several independent genes can lead to similar clinical features, implying both phenotypic and genetic overlaps (Reiter and Leroux, 2017). Although various observational studies and case series have been published before, most of them focused on the genetic rather than the phenotypic presentation (König et al., 2017). It is thus indispensable to have subgroups of patients identified by similar phenotypes, and combine with genetic subgroups to achieve better stratification.

With the large volume of clinical data being collected, phenotyping from these data, including Electronic Health Records (EHRs) and clinical research data, has been recognized as the basic staple to enable precise diagnoses, subtyping, and treatment (Weng et al., 2018). EHRs are a rich source of phenotype information that can be in various formats, including coded data (e.g. ICD codes), numerical measurements (e.g. laboratory test results), images, and more importantly, unstructured narrative/textual data (Frey et al., 2014; Sharafoddini et al., 2017), which contain less standardized information like early and unexplained signs and symptoms, history of the disease, diagnostic hypotheses, and so forth. Although combination of multiple modalities of EHRs can achieve better performance (Zeng et al., 2019), EHR data are collected during patients' healthcare for the purpose of delivering medical treatment rather than phenotyping. Consequently, they do not have the same consistency and precision of data collected for experiments (Frey et al., 2014). Another main source of phenotyping is clinical and scientific research data, which are collected toward specific goals. Research data thus provide standardized and in-depth information of targeted cohort of patients for deep phenotyping, especially in the situation of rare disease, where the knowledge about the disease is limited and keeps evolving.

In this article, we describe the approaches that we have developed for (1) deep phenotyping from different modalities of EHR data (narratives and quantitative data) and also from research data, (2) defining patient phenotypic similarity from clinical data warehouse with the objective of helping fast and precise diagnosis and subtyping in the context of complex rare disease. The methods were applied to overcome the medical challenge of identifying patients with ciliopathies, and stratifying them into subgroups with respect to their phenotypic characteristics. This study was conducted as part of the C'IL-LICO program, which aims at developing transformative diagnostic, prognostic, and therapeutic approaches for patients suffering from ciliopathies.

2. Materials and methods

2.1. Material

As a French national reference center for rare and undiagnosed diseases, the Necker-Enfants Malades Hospital (Necker Children's Hospital) hosts the Imagine Institute, a research institute focusing on genetic diseases. The clinical data warehouse (Dr. Warehouse®) contains on one hand clinical and scientific research data, and on the other hand, EHR data, which makes a total of 696 401 patients and 5 832 165 documents (Supplementary Appendix 1). Such data warehouse is to our knowledge the biggest data repository of rare disease France. The high throughput phenotyping within the Dr. Warehouse® system is based on the extraction of Unified Medical Language System (UMLS) concepts (Supplementary Appendix 1).

The data repository of Necker/Imagine contains more than 1300 patients with known classified diseases or syndrome associated with ciliopathy. Among them, the nephronophthisis (NPH) cohort is one of the major ciliopathy cohorts (Supplementary Appendix 2). The objectives of the C'IL-LICO project include (i) detection of undiagnosed cases of ciliopathy that would benefit from genetic testing (Figure 1A), and (ii) stratification of diagnosed ciliopathy patients (Figure 1B). More details about the two tasks and the evaluation schema will be presented later in section 2.4. Our approach is based on firstly, deep phenotyping of ciliopathies and, secondly, computing subjacent similarities between patients followed up at Necker/ Imagine.

2.2. Deep phenotyping of ciliopathies

2.2.1. Phenotyping from research data (NPH dataset)

As a reference center in France, patients diagnosed or suspected to suffering from a NPH-related ciliopathy are referred to Necker/Imagine for research and follow up. A comprehensive questionnaire is completed for all index patients and their family members. This questionnaire has been designed by experts to reflect all the current knowledge on ciliopathy and support precise and standardized phenotyping. We will refer the questionnaire data of NPH-related ciliopathies as NPH dataset in the rest of this article. The data were processed as following:

- Each term in the questionnaire was mapped to the UMLS.
- The free-text comments in the questionnaire were processed as for the EHR, using the method developed in Dr. Warehouse®, i.e. all the phenotype concepts are extracted from text as UMLS concepts.

2.2.2. Phenotyping from EHRs with disambiguation

For those patients followed up at Necker Children's Hospital, their EHRs are available in Dr. Warehouse®. As described in (Garcelon et al., 2018), the concept extraction module in Dr. Warehouse® extracts all phenotype information from the EHR and distinguishes between (i) the phenotypes of the index patient and their family history, (ii) the different modalities of a term, like negation. In this study, we considered only patient's own "positive" concepts to build the data matrix for computing the similarity.

Besides negation and family history, we identified two additional sources of false positives that required developing disambiguation algorithms. The first category occurs when the terms used for a laboratory test and its results may be identical, for example, "proteinuria" as lab test and "proteinuria" as phenotype. To address this issue, we used the corresponding quantitative laboratory test results to distinguish between the annotation by one semantic type or the other. The normal threshold of related lab tests with all possible measurements and units have been defined. As for proteinuria, possible measurements include: proteins in urine (g/l), proteins in 24-hour urine collection (g/24H), and protein-creatinine ratio on spot urine sample (mg/mmol). Patient's phenotype is defined as presence of sign (e.g., proteinuria), if at least one lab test result is abnormal. Based on this method, all false positive extractions were removed from the database.

Another source of false positives was the list of short terms for vaccines, e.g., hepatitis B, measles. This was addressed by defining a pattern of vaccination as the presence of at least two concepts together of such list of diseases, like {measles, rubella, mumps}. All related extractions were excluded from the patient's phenotypes.

The result of the whole phenotyping step is a set of UMLS concepts and their provenance metadata: (1) ciliopathy questionnaire or (2) regular EHR.

2.3. Methods of similarity

In order to compute the similarity metrics, we considered patient representation as the set of all his/her phenotype UMLS concepts, thus extractions from family history and negative extractions were excluded. For patient i, we considered two formats of representation, (i) binary data, $x_{ij} = 1$ if patient i has phenotype j, and $x_{ij} = 0$ if patient i doesn't have phenotype j; (ii) frequence data, x_{ij} is the number of occurrences of phenotype j in all documents of patient i. For each format, three normalizations were computed: by column (distribution of phenotypes over patients), by row (distribution of patients over phenotypes),

and by term frequency–inverse document frequency (tf-idf). The normalization by column will penalize the presence of a common phenotype in the similarity measure. For example, concepts like “fever”, “pain” and “asthenia” can be very frequent in EHR, but providing less interesting information. By dividing the column sum (the number of patients with the phenotype), the contribution of such common concept in the similarity measure is reduced. The normalization by row will adjust the imbalance of number of phenotypes between different patients. Actually, some patients in our database have a long-term follow-up (max 28.6 years); moreover if the diseases are complicated affecting multiple organs, one patient may have more than 500 distinct concepts in EHR. With a dot product similarity metric, these patients have more chance to be similar with everyone else. By dividing the row sum (the number of phenotypes that a patient has), such issue can be addressed. The tf-idf is largely used in information retrieval to reflect how important a word is to a document in a collection or corpus. In our case, the tf-idf of patient i for phenotype j, reflecting how important phenotype j is to patient i in the cohort, was obtained with the formula $(x_{ij}/\sum_k x_{ik})*\log(N/\sum_h \mathbb{I}_{x_{hj}>0})$, where N is the total number of patients, and the sum in the log is the number of patients with the concept j. It takes into account both aspects discussed above.

Cosine similarity, which is defined as $\sum_k x_{ik}x_{i'k}/(\sqrt{\sum_k x_{ik}^2}\sqrt{\sum_k x_{i'k}^2})$ for two vector $\boldsymbol{x_i}$ and $\boldsymbol{x_{i'}}$, is largely used in the context of text/document. In our case, the number of phenotypes per patient can vary greatly, depending on many factors, including large spectrum of clinical features from one affected organ to multisystemic defects, as well as the time of follow-up, the doctors’ habit, etc. Two patients with very different number of phenotypes can be still considered as similar if they share very few but characteristic phenotypes. Therefore, here we consider the dot product similarity between patient i and patient i', $\sum_k x_{ik}x_{i'k}$, which is based only on the phenotypes in common.

2.4. Evaluation and application

In order to evaluate the computed similarity metrics, we removed from the set of each patient’s UMLS concepts:

- all the concepts also belonging to the UMLS semantic type ‘Gene or Genome’
- the concepts corresponding to the diagnosis of the patient, e.g., nephronophthisis.

We considered two tasks: (1) identifying NPH-related ciliopathies from other nephropathies, (2) subtyping diagnosed NPH-related ciliopathies. The evaluation schema is shown in Figure 1. For the first task, we considered on one hand the EHRs of the NPH patients, and, on the other hand the EHRs of the patients with 'other nephropathy', i.e. any kidney disease (CUI C0022658 and all its descendants) excluding ciliopathy. As renal involvement is one of the most frequent manifestations in ciliopathies, the hypothesis is that the cohort of patients with any symptom related with kidney disease may contain new patients suspected to suffer from a ciliopathy. The EHR datasets were pooled to compute similarities. The ranking list of patients will be provided with the average similarity with diagnosed ciliopathy patients, where the top ranking patients are those most similar with ciliopathies. By moving downward the similarity threshold, the precision (percentage of relevant patients out of all predicted patients), defined as $|\{\text{diagnosed patients}\} \cap \{\text{predicted patients}\}| / |\{\text{predicted patients}\}|$, decreases, while the recall (percentage of predicted relevant patients of all relevant patients), defined as $|\{\text{diagnosed patients}\} \cap \{\text{predicted patients}\}| / |\{\text{diagnosed patients}\}|$, increases. When making a predictive decision with a fixed threshold k, i.e. predicting the top k patients in the ranking list (most similar with ciliopathies) as patients suspected to suffer from a ciliopathy, the precision@k, defined as $\{\text{diagnosed patients in the top } k\}/k$, can be considered for the performance of this predictive decision. The area under the precision-recall curve (AUPR) will be also considered, as it summarizes the trade-off between true positive rate and the positive predictive value for our model with different thresholds.

For the second task (subtyping), as more published studies focused on the genetic rather then phenotypic presentation, we aim at stratifying patients into subgroups with respect to their phenotypic characteristics extracted from the NPH dataset. As presented above in the introduction, ciliopathies are pleiotropic, i.e. mutations of the same disease-causing gene can result in distinct clinical entities, and, conversely, mutations in several independent genes can lead to similar clinical features, we considered a comparison between phenotypic and genotypic information. We thus evaluate the internal and external phenotypic similarities of each mutated gene group. A spectral clustering of patients with similarity matrix will be performed and compared to the gene labels.

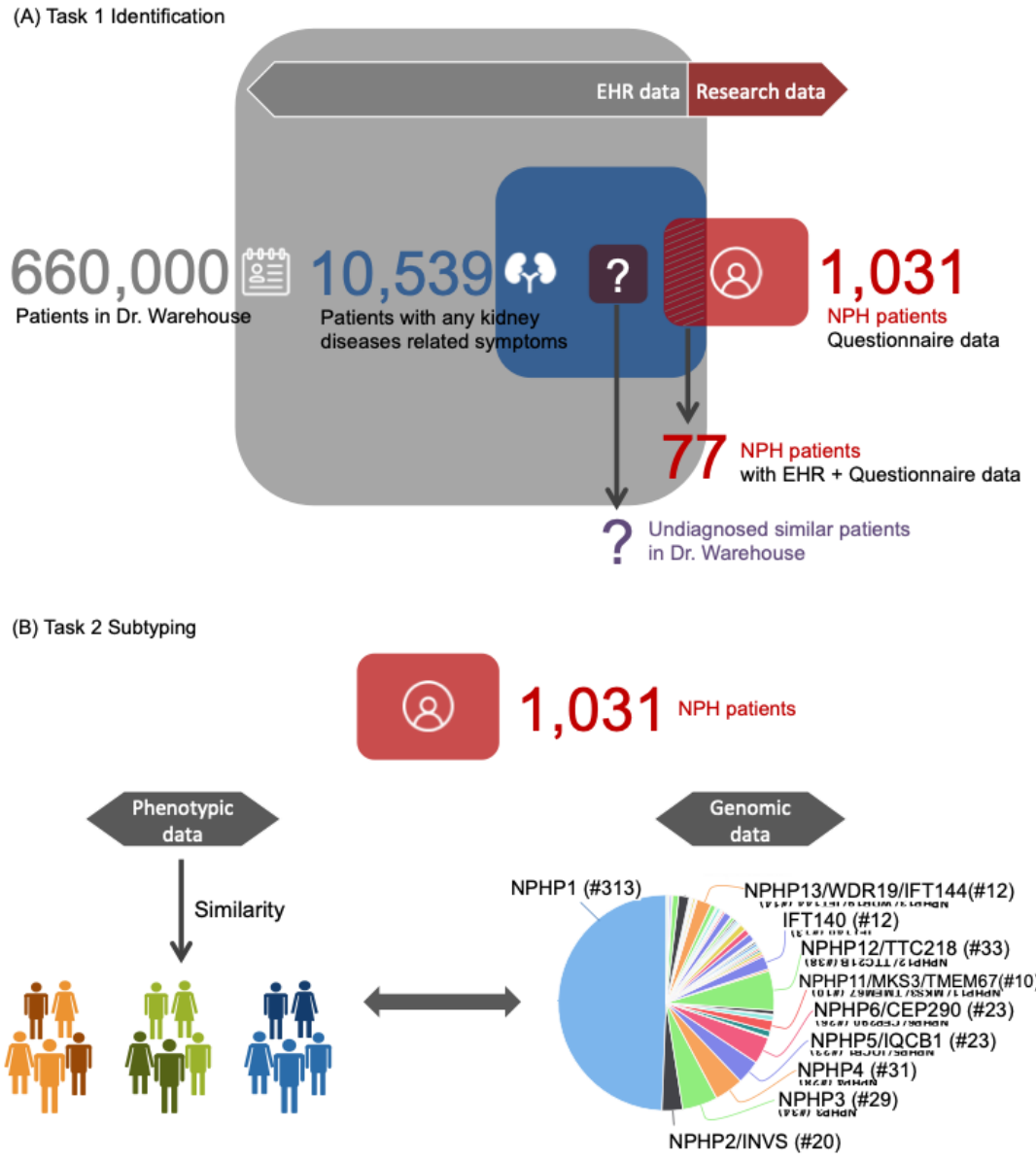

Figure 1 Schema of evaluation

3. Results

3.1. Dataset

We obtained 1,031 diagnosed patients in the NPH dataset， all with questionnaire phenotypic data. Among them, 633 patients were found with at least one mutation homozygous of screened ciliopathy related genes. These diagnosed patients have in average 5.9 phenotypes (sd=3.8) described in the questionnaire. The most frequent semantic type is ‘Disease or Syndrome’, followed by ‘Finding’, ‘Congenital Abnormality’ and ‘Sign or Symptom’ (Table 1). The most frequent concepts include Nephropathy, Renal insufficiency, End-stage

renal disease, Polyuria-Polydipsia, Congenital anomaly of eye, Morphological abnormality of the central nervous system, Abnormality of the liver, Reduced visual acuity, Hypertension, Intellectual disability (Figure 2).

Based on Dr. Warehouse®, 77 of the diagnosed patients in NPH database were followed up at Necker Children's Hospital with their EHR data available, which contained in average 47.7 phenotypes (sd=46.3) per patient. The most present semantic type is 'Disease or Syndrome', followed by 'Finding', 'Pathologic Function' and 'Sign or Symptom' (Table 1). The ten most frequent concepts are: Nephropathy, End-stage renal disease, Renal insufficiency, Systemic blood pressure, Proteinuria, Cyst, Hypertension, Anastomosis, Pulse, Fever and Anemia. These patients have an average follow-up of 6.5 years (range [0.7, 28.6]).

Noticeably, the granularity and the coverage of the data vary from one data source to the other (Table 1 and Figure 2). There are more concepts in EHRs than in questionnaire for all semantic categories, but the distribution is slightly different, i.e. congenital abnormalities are more addressed in questionnaire. The information is more concise, refined and targeted to ciliopathy in NPH dataset, while EHRs contain more general symptoms like Fever, Anemia and Fatigue, which may be the symptoms that motivated the consultation or appeared during follow-up. The concepts corresponding to Anastomosis and Pulse are present in the EHR when patients are treated by dialysis, which are related to renal insufficiency.

Semantic type	NPH dataset (questionnaire)			EHR_Ciliopathy			EHR_Nephropathy		
	# concepts in total	% concepts in total	# concept per patient	# concepts in total	% concepts in total	# concept per patient	# concepts in total	% concepts in total	# concept per patient
Disease or Syndrome	168	50%	3.5	818	53%	20.9	4377	50%	23.6
Congenital Abnormality	54	16%	0.5	70	5%	1.1	447	5%	1.1
Finding	48	14%	0.8	323	21%	10.2	1574	18%	10.7
Sign or Symptom	28	8%	0.5	201	13%	6.0	697	8%	6.6
Pathologic Function	25	8%	0.1	161	10%	6.7	633	7%	7.5
Anatomical Abnormality	14	4%	0.0	41	3%	1.2	217	2%	1.3
Neoplastic Process	11	3%	0.0	64	4%	1.2	663	8%	1.7
Mental or Behavioral Dysfunction	9	3%	0.1	56	4%	0.7	197	2%	0.7
Acquired Abnormality	0	0%	0	23	1%	0.8	150	2%	0.8
Physiologic Function	0	0%	0	20	1%	1.0	56	1%	0.8

Table 1 Category of UMLS concepts extracted from questionnaire of NPH dataset and EHRs (EHRs of patients with diagnosed ciliopathies and EHRs of patients with other nephropathies are shown separately). The number and percentage of concepts in each of ten semantic categories are shown for three data sources, as well as the number of concepts per patient in each semantic category.

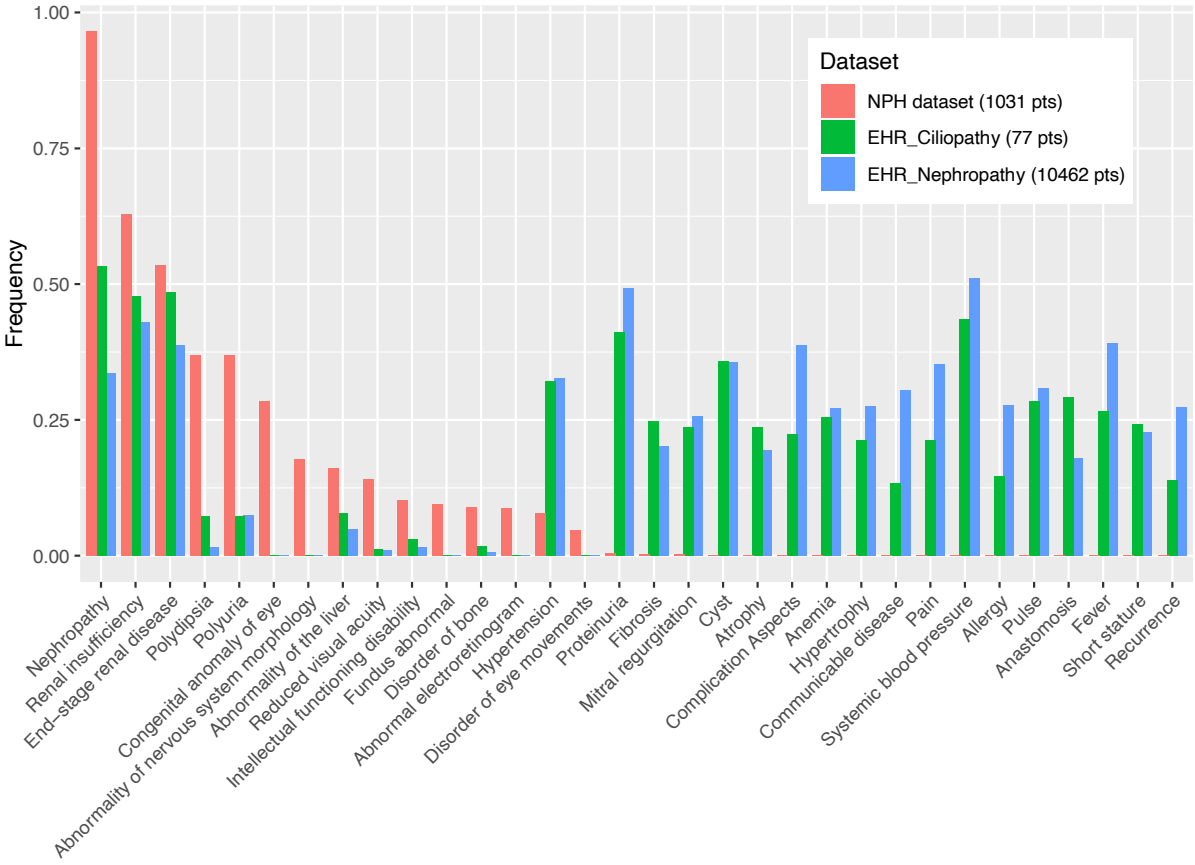

Figure 2 Frequent concepts distributions of NPH dataset, EHR of ciliopathies and EHR of other nephropathies.

Regarding the first task (identification), we extracted 10,539 patients from Dr. Warehouse® with any symptom related to kidney diseases, including the 77 diagnosed ciliopathies patients, and 10,462 patients with 'other nephropathy'. EHRs of 'other nephropathy' patients exhibited no significant difference with the EHRs of the 77 NPH patients in terms of average number of concepts per patient, semantic types of the concepts and the distribution of the most common concepts (Table 1 and Figure 2). We refer this dataset as Evaluation Set 1. The diagnostic terms corresponding to Ciliopathies (C4277690), Nephronophthisis (C0687120), Jeune thoracic dystrophy (C0265275), Alstrom syndrome (C0268425), Joubert syndrome (C0431399) and Bardet-Biedl syndrome (C0752166) were removed for the 77 NPH patients. The terms belonging to the UMLS semantic type 'Gene or Genome' were removed as well. The resulting Evaluation Set 1 consists thus 10,539 patients, presenting 8,738 phenotypes with distinct UMLS code.

For the second task (subtyping), we considered the 1,031 patients in NPH dataset, presenting 333 phenotypes with distinct UMLS code. We integrated genetic data with phenotypic data for all 1,031 patients, and refer this dataset as Evaluation Set 2. Predictive damaging variants were found in 62 different genes. NPHP1 homozygous deletion was the

most represented group, with 313 patients, followed by *NPHP12/TTC21B, NPHP4, NPHP3, NPHP5/IQCB1, NPHP6/CEP290, NPHP2/INVS, NPHP13/WDR19-IFT144, NPHP8/MKS3/RPGRIP1L, IFT140* and *NPHP11/MKS3/TMEM67*, with 10 to 35 patients (Figure 1B). For the remaining groups, most of them contain less than two patients. We thus focus on these 11 gene groups with more than 10 patients mutated.

3.2. Results for task 1: identifying ciliopathies from other nephropathies

Both binary data and frequency data were tested with three normalizations: by column (distribution of phenotypes over patients), by row (distribution of patients over phenotypes), and by tf-idf. Table 1 and Figure 3A summarized the results of task 1. In general, the frequency data outperformed binary data in terms of precision@k and AUPR. The normalization by column returned the best precision@k for k=30, k=50 and k=100. For k=30, we obtained a precision of 0.767, which means that in the top 30 ranking patients, 23 of them are diagnosed patients. When increasing k to 50, the best result returned 0.56 of precision, corresponding to 28 diagnosed ciliopathies; while in the top 100 ranking patients, 35 of them are diagnosed patients, corresponding to nearly half of all ciliopathy patients in Evaluation Set 1.

Task 1	Precision@k			AUPR
	k=30	k=50	k=100	
bi_norm1	0.733	0.480	0.330	0.393
bi_norm2	0.160	0.155	0.095	0.039
bi_tfidf	0.633	0.460	0.310	0.372
freq_norm1	**0.767**	**0.560**	**0.350**	**0.399**
freq_norm2	0.067	0.057	0.047	0.026
freq_tfidf	0.600	0.440	0.310	0.330

Table 2 Performance of task 1 (identification). Similarities are computed for both binary and frequency data with three normalizations. bi_norm1: binary data normalized by column; bi_norm2 : binary data normalized by row; bi_tfidf : binary data normalized by tf-idf; freq_norm1 : frequency data normalized by column; freq_norm2: frequency data normalized by row; freq_tfidf: frequency data normalized by tf-idf.

In order to better understand the EHR data, we investigated whether the results can be improved by using only characteristic phenotypes and ignoring noisy phenotypes for similarity. We further applied three variable selection algorithms (linear support vector selection with $l2$ penalty, tree-based selection and random forest selection), and took only the selected relevant variables (phenotypes) to compute similarity. The three selection

models returned respectively 1201, 704 and 601 relevant variables out of 8738 phenotypes with default settings. However, the improvement was not significant (Figure 3B).

Our result showed that the ranking list of similarity could provide reliable diagnosis when we restrict to the very top. This can be partially explained by the phenotypic variability of the ciliopathies, i.e. patients presenting with multisystemic symptoms are easier to identify from other nephropathies than patients presenting with isolated renal symptoms.

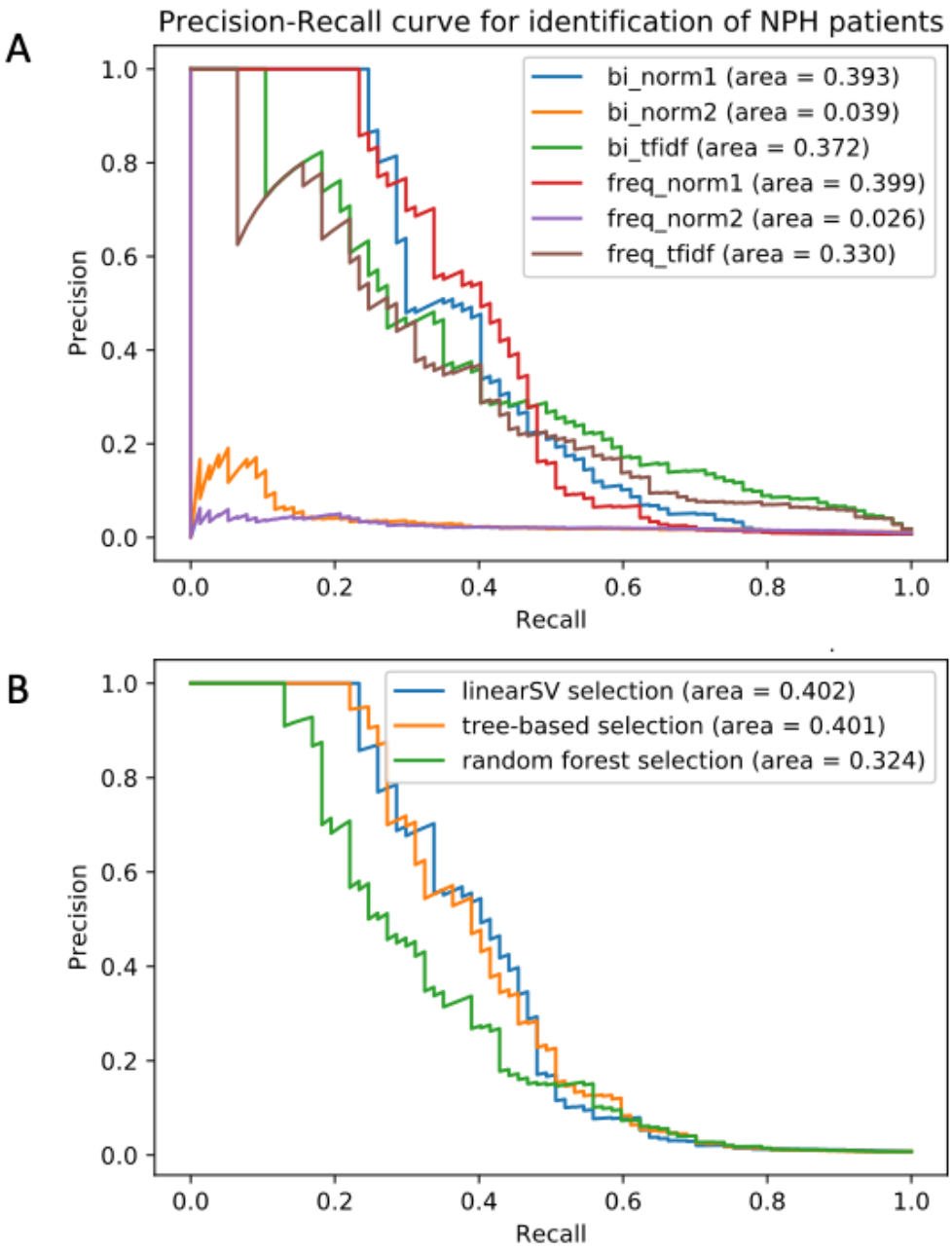

Figure 3 Precision-Recall curves for identification of NPH patients. A: Binary data and frequency data with three normalizations: by column, by row and by tfidf. B: Three variable selection models. bi_norm1: binary data normalized by column; bi_norm2 : binary data normalized by row; bi_tfidf : binary data normalized by tf-idf; freq_norm1 : frequency data normalized by column; freq_norm2: frequency data normalized by row; freq_tfidf: frequency data normalized by tf-idf; linearSV: linear support vector.

3.3. Results for task 2: subtyping diagnosed ciliopathies

Although EHR data may contain early signs helpful for diagnosis, once the patient is diagnosed, the specific questionnaire data suits better for the subtyping task. Therefore, we focused on questionnaire data for this task. Similarities based on binary data were used for spectral clustering, where the number of clusters was fixed to 11, which equals to the number of genes mutated in more than 10 patients in our NPH dataset. Meanwhile it is also a reasonable number for qualitative comparison between clinical clusters and mutated gene groups. Various evaluation metrics, including silhouette score, homogeneity, completeness and V-measure are considered for the comparison. The binary data normalized by tf-idf (with all non-zero $x_{ij} = 1$) provided the best result (Table 2).

We further computed the internal-external similarities for the most common mutated gene groups. Figure 4A showed the results with a heatmap, where the red color presented higher similarity while the yellow color presented lower similarity. We observed on the diagonal internal similarities: most of the mutated gene groups showed a high internal similarity except NPHP1 and NPHP4, the latter could be explained by the large phenotypic variability in these groups of patients. In these two cases, the majority of the patients presented with isolated renal symptoms (with less number of phenotypes), whereas the other patients presented additionally retinal degeneration or neurologic defects (Senior-Løken syndrome, Cogan syndrome, Joubert syndrome). On the off diagonal, we observed external similarities: several regions of strong associations were marked, which were consistent with expert knowledge (Figure 4B). We provided here short explanations for each discussion point marked on the heatmap. (1) High similarity was observed between NPHP5 and NPHP6 groups. The protein products of these two genes are known to form a specific functional complex at the connecting cilium of photoreceptors (Sang et al., 2011). All patients with either NPHP5 or NPHP6 mutations present with eye abnormalities, always associated with progressive blindness (Senior-Loken syndrome, SLS) (Mitchison and Valente, 2017), explaining the high similarity. (2) NPHP6 patient group was also similar to NPHP8 patient group, as mutations in these two genes, encoding both transition zone proteins, can be associated with a multisystemic disease with brain stem malformation (Joubert syndrome, JBTS) and extensive embryonal defects (Meckel-Guber syndrome, MKS) (Wolf et al., 2007). (3) Strong

external similarities were also observed between NPHP12, NPHP13 and IFT140 groups, which are in fact components of the intraflagellar transport complex A (IFT-A) that drives retrograde ciliary transport required for ciliary signaling and maintenance. Mutations of IFT-A genes generally cause cilia-related bone phenotypes (Jeune (JATD), Sensenbrenner (CED), Saldino-Mainzer (SDMZ) syndromes), renal defects (nephronopthisis type), as well as liver and retinal anomalies. (4) The NPHP3 and NPHP11 patient groups presented high intrinsic similarities, and also showed high external similarities with the IFT-A groups, which can be partly explained by the strong associated phenotypes: liver, kidney and retinal affections. Accordingly, functional interaction has been reported between IFT-A and transition zone proteins.

Task 2	Silhouette	Homogeneity	Completeness	V-measure
bi	0.072	0.392	0.228	0.288
bi_norm1	0.004	0.323	**0.261**	0.289
bi_norm2	0.140	0.376	0.241	0.294
bi_tfidf	**0.226**	**0.415**	0.255	**0.316**

Table 3 Performance of task 2 (subtyping). Similarities are computed for binary data with three normalizations. bi: binary data without normalization; bi_norm1: binary data normalized by column; bi_norm2 : binary data normalized by row; bi_tfidf : binary data normalized by tf-idf.

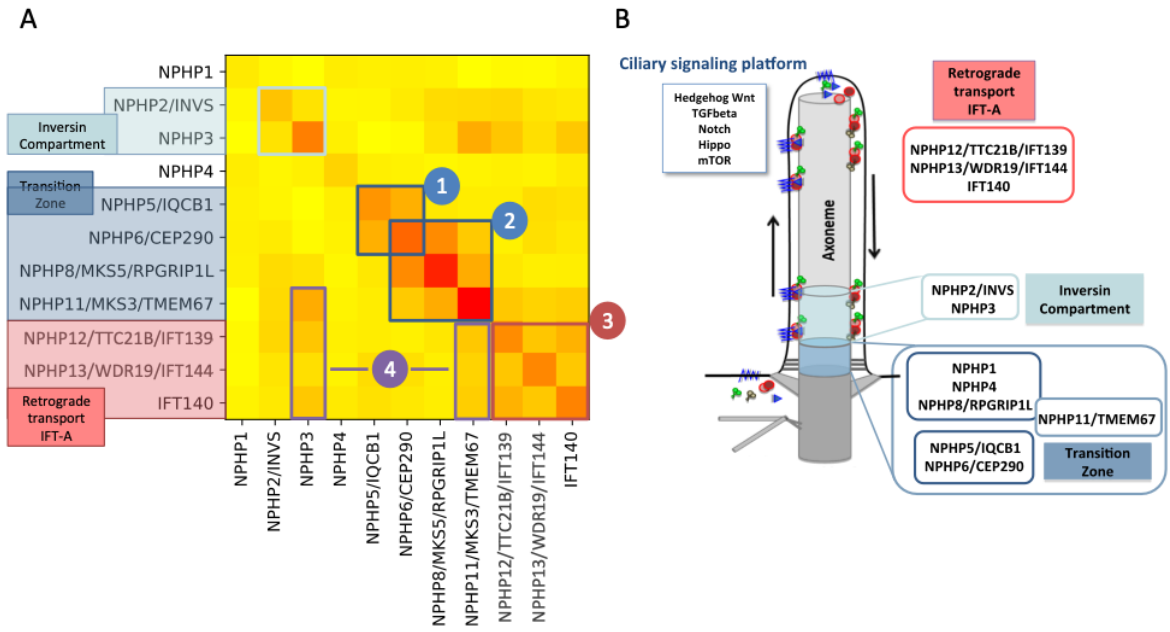

Figure 4 Association of main mutated genes. A: Internal-external similarities of main patient groups. In the heatmap, the red color presents higher similarity while the yellow color presents lower similarity. Four regions of strong associations are marked. B: Ciliary signaling mechanism. NPH-related genes encode proteins present at the primary cilium and forming

functional networks. The transition zone, inversin compartment and IFT proteins are involved in intra-flagellar transport. The corresponding genes are marked in the left panel A with the same colors.

4. Discussion

We introduced in this paper a new pipeline of using deep phenotyping for defining patient similarity in order to help fast and precise diagnoses and subtyping. Our study pursues the previously published method for identifying rare disease patients with EHR data, which was developed in Necker Children's Hospital (Garcelon et al., 2017). This method used for deep phenotyping consists in extracting comprehensive and fine-grained descriptions of patients' clinical phenotypes, and mapping them to UMLS. It was evaluated for simple rare diseases. Here we improved the phenotyping with disambiguation by combining multiple modalities of clinical data (narratives and quantitative data) for complex pleiotropic rare diseases. This type of complex rare disease requires as well precision medicine approach to identify clinical meaningful subtypes. We therefore extended the method of similarity for subtyping purpose by using multiple data sources (clinical research data and EHRs). Our preliminary results in ciliopathies can be summarized as (1) the ranking list of patients similar with ciliopathies could provide reliable diagnosis predictions; (2) the phenotypic similarity among ciliopathy patients showed strong concordance with expert knowledge, thus, together with in-depth genomic similarity, it is promising to define better subtyping of ciliopathy.

4.1. Comparison to other work

According to a recent systematic review of 279 articles published between 2012 and 2017 (Parimbelli et al., 2018), the most represented category of data type used for patient similarity is molecular data, followed by clinical data and imaging/biosignals data, and the most frequently considered clinical domain is cancer, followed by nervous system, integumentary system, respiratory system and digestive system. Ow et al. used Euclidean distance and Kendall's Tau rank correlation on expression data of predefined mRNA predictors to stratify patients of high-grade serous ovarian cancer into prognostic subgroups (Ow et al., 2016). Wang et al. proposed a semi-supervised recursive tree partitioning approach by using radial basis function (RBF) kernel as similarity metric, which is exponential of negative squared Euclidean distances, and evaluated on breast cancer datasets, where each patient is represented by 30 features of breast mass image (Wang, 2015). Regarding patients' phenotype, secondary use of EHR data has been considered a key solution for phenotyping. However, in clinical settings, EHR data are collected for the purpose of delivering medical treatment at the point of care rather than phenotyping, they do not have the same consistency and precision of data collected for experiments (Frey et al.,

2014). The challenges such as incompleteness, inaccuracy and complexity have been much discussed (Hripcsak and Albers, 2013). Many studies of deep phenotyping focused on a set of phenotypic features defined based on expert knowledge of specific diseases or medical issues (Kopf et al., 2018; Peron et al., 2018; Radley et al., 2019), therefore, data were collected for a specific scientific and clinical research purpose to ensure the precision and comprehensiveness of phenotyping. Recently, Zhang et al. presented their method for mapping LOINC-encoded laboratory test results transmitted in Fast Healthcare Interoperability Resource (FHIR) standards to Human Phenotype Ontology (HPO) terms (Zhang et al., 2019), which shared the same idea as we have for ciliopathy, i.e. using thresholds on quantitative lab test to transform them into phenotypes. The advantage of our method is that we have gathered all phenotypes from all data sources of all patients, and we have defined more appropriate categories instead of using standard thresholds to confirm or reject related phenotype extraction. For example, as low level of proteinuria is observed in patients with mutations in NPH genes, we defined with nephrologists more adapted categories to distinguish nephrotic syndromes and "mild proteinuria" in order to achieve precise phenotyping. With phenotype data, Zhang et al. used Jaccard index to compute similarities between patient vectors, where each patient is a binary vector of ICD9 diagnosis categories, and the Jaccard index is in fact the number of ICD9 codes in intersection divided by the number of ICD9 codes in union of two patients (Zhang et al., 2014). There are few works in the context of rare disease, Greene et al. used semantic similarity (Lin's similarity function) to compare two different HPO terms, then aggregated for a set of HPO terms (representation of a patient) by averaging Lin's similarities with best match term (Greene et al., 2016). In our study, the EHRs are in French language, thus the annotation is achieved with UMLS French (SAP=FRE) in our data warehouse. We did not use HPO because it does not exist yet a reliable French version. We considered only patient's own "positive" concepts to define similarity; therefore, "negative" concepts (like "patient doesn't have fever") and concepts extracted from family history were excluded. As similarity is defined on the presence of clinical phenotype feature, Euclidean based metrics is less suitable than vector space model, and our experiences have also confirmed this statement. As mentioned previously, the number of phenotypes per patient can vary greatly, depending on many factors, including large spectrum of clinical features from one affected organ to multisystemic defects, as well as the time of follow-up, the doctors' habit, etc., thus two patients with very different number of phenotypes can be still considered as similar if they share very few but characteristic phenotypes. Therefore, here we focused on the phenotypes in common, and didn't use any denominator as for Jaccard index or cosine similarity. In this work, we didn't use a semantic similarity, because UMLS network is less formally structured than HPO network, thus the semantic similarity between two concepts based on the number of descendant of the lowest common ancestor gave less reliable results.

4.2. Strengths and limitations

We integrated both dedicated questionnaire data and EHR using the UMLS as a core thesaurus for phenotyping. The UMLS provides a set of terms, concepts, and semantic types that was used for information extraction, semantic integration, and categorization. Moreover, we integrated several data types, including phenotypes extracted from narrative reports and structured data like lab test results. We showed that multi-source phenotyping could be used in disambiguation tasks to reduce false positives and improve precision. More precisely, integration of quantitative lab test results enabled identification of several false positives (e.g., proteinuria) generated by the NLP module. In fact phenotyping was performed in two steps: first, all the phenotypes were extracted without removing such false positives, then the false positives were removed by comparing with the quantitative lab test results. We tested the similarity metrics before and after removing such false positives, and demonstrated that in task 1 (identification), the precision@30 improved from 0.73 to 0.77. Our conclusion is that data preprocessing, cleaning, and phenotyping approaches are crucial steps in any study re-using EHR data, and that all phenotyping steps must be validated to guarantee that the models learned from the data are unbiased. This conclusion is shared by other authors (Denaxas et al., 2017) (Williams et al., 2017) (Yu et al., 2015).

The extraction module of Dr. Warehouse® may produce false negatives because of its exact match strategy. However we did not perform an extensive review of the data to estimate the recall performance (absence of false negatives) of the phenotype extraction step in ciliopathies. Moreover, phenotype concepts have many features (e.g., the anatomical site) that can be modeled in formal ontologies and used to establish their dependencies with other concepts. For example, an extraction of ‘cyst’ can be related more accurately to ‘renal cyst’ or ‘hepatic cyst’. However in this work we did not explore further this issue.

The longitudinality of phenotype has not been taken into account for computing similarity in this work, which raises an important issue for ciliopathy, i.e. a patient presenting with eye abnormalities in infancy then renal affection can be very different with a patient showing renal disorders first and progressive eye abnormality in adulthood.

4.3. Conclusion

Most research on similarity in the medical domain aim at clustering patients in more homogeneous subgroups that could explain different outcomes for what was considered before the “same” disease, for example in oncology. Interestingly, in 2011, Frankovich et al. reported on the use of their clinical data warehouse to make therapeutic decision about a young patient with systemic lupus erythematosus (Frankovich et al., 2011). More precisely,

since there were neither guidelines nor consensus on the decision, they decided to query their clinical data warehouse to review similar cases and made decision on the basis of the results. This example illustrates new scenarios based on searching for similar patients in a data warehouse that may be of great help in the domain of rare diseases. The objective of C'IL-LICO includes both a better classification of ciliary dysfunction and providing a similarity-based mechanism to detect patients that "look similar to" ciliopathy patients. Similarity metrics applied to rare disease offer new perspectives in a translational context that may help to recruit patients for research and reduce the length of the diagnostic journey. In our case of rare complex disease - ciliopathy, similarity metrics based on deep phenotyping enable better subtyping from both clinical and genetic characteristics to achieve precision medicine.

Ethics approval

We got an ethical approval by the French IRB CPP Il-de-France II (IRB registration number 00001072) registered under reference 2016–06-01 for the clinical data warehouse. And for the nephronophthisis database, we got for Clinical protocol: CPP Il-de-France II approval on 04/07/2016 and ANSM authorization on 17/06/2016; Biological collection: approval of CPP Il-de-France II on 04/05/2015 and of the Advisory Committee on Information Processing in Material Research in the Field of Health (CCTIRS) on 11/02/2016.

Acknowledgement

This work was supported by State funding from The French National Research Agency (ANR) under "Inverstissements d'Avenir" programs (Reference: ANR-10-IAHU-01) and C'IL-LICO project (Reference: ANR-17-RHUS-0002).

Competing interests

The authors declare that they have no competing interests.

References

Blöß, S., Klemann, C., Rother, A.-K., Mehmecke, S., Schumacher, U., Mücke, U., Mücke, M., Stieber, C., Klawonn, F., Kortum, X., Lechner, W., Grigull, L., 2017. Diagnostic needs for rare diseases and shared prediagnostic phenomena: Results of a German-wide expert Delphi survey. PLoS ONE 12. https://doi.org/10.1371/journal.pone.0172532

Denaxas, S., Direk, K., Gonzalez-Izquierdo, A., Pikoula, M., Cakiroglu, A., Moore, J., Hemingway, H., Smeeth, L., 2017. Methods for enhancing the reproducibility of biomedical research findings using electronic health records. BioData Min. 10, 31. https://doi.org/10.1186/s13040-017-0151-7

Dharssi, S., Wong-Rieger, D., Harold, M., Terry, S., 2017. Review of 11 national policies for

rare diseases in the context of key patient needs. Orphanet J. Rare Dis. 12. https://doi.org/10.1186/s13023-017-0618-0

Franco, P., 2013. Orphan drugs: the regulatory environment. Drug Discov. Today 18, 163–172. https://doi.org/10.1016/j.drudis.2012.08.009

Frankovich, J., Longhurst, C.A., Sutherland, S.M., 2011. Evidence-Based Medicine in the EMR Era. N. Engl. J. Med. 365, 1758–1759. https://doi.org/10.1056/NEJMp1108726

Frey, L.J., Lenert, L., Lopez-Campos, G., 2014. EHR Big Data Deep Phenotyping. Contribution of the IMIA Genomic Medicine Working Group. Yearb. Med. Inform. 9, 206–211. https://doi.org/10.15265/IY-2014-0006

Garcelon, N., Neuraz, A., Benoit, V., Salomon, R., Kracker, S., Suarez, F., Bahi-Buisson, N., Hadj-Rabia, S., Fischer, A., Munnich, A., Burgun, A., 2017. Finding patients using similarity measures in a rare diseases-oriented clinical data warehouse: Dr. Warehouse and the needle in the needle stack. J. Biomed. Inform. 73, 51–61. https://doi.org/10.1016/j.jbi.2017.07.016

Garcelon, N., Neuraz, A., Salomon, R., Faour, H., Benoit, V., Delapalme, A., Munnich, A., Burgun, A., Rance, B., 2018. A clinician friendly data warehouse oriented toward narrative reports: Dr. Warehouse. J. Biomed. Inform. 80, 52–63. https://doi.org/10.1016/j.jbi.2018.02.019

Global rare disease commission, 2018. Ending the Diagnostic Odyssey for Children with a Rare Disease.

Greene, D., Richardson, S., Turro, E., 2016. Phenotype Similarity Regression for Identifying the Genetic Determinants of Rare Diseases. Am. J. Hum. Genet. 98, 490–499. https://doi.org/10.1016/j.ajhg.2016.01.008

Hripcsak, G., Albers, D.J., 2013. Next-generation phenotyping of electronic health records. J. Am. Med. Inform. Assoc. JAMIA 20, 117–121. https://doi.org/10.1136/amiajnl-2012-001145

König, J., Kranz, B., König, S., Schlingmann, K.P., Titieni, A., Tönshoff, B., Habbig, S., Pape, L., Häffner, K., Hansen, M., Büscher, A., Bald, M., Billing, H., Schild, R., Walden, U., Hampel, T., Staude, H., Riedl, M., Gretz, N., Lablans, M., Bergmann, C., Hildebrandt, F., Omran, H., Konrad, M., Gesellschaft für Pädiatrische Nephrologie (GPN), 2017. Phenotypic Spectrum of Children with Nephronophthisis and Related Ciliopathies. Clin. J. Am. Soc. Nephrol. CJASN 12, 1974–1983. https://doi.org/10.2215/CJN.01280217

Kopf, S., Groener, J.B., Kender, Z., Fleming, T., Bischoff, S., Jende, J., Schumann, C., Ries, S., Bendszus, M., Schuh-Hofer, S., Treede, R.-D., Nawroth, P.P., 2018. Deep phenotyping neuropathy: An underestimated complication in patients with pre-diabetes and type 2 diabetes associated with albuminuria. Diabetes Res. Clin. Pract. 146, 191–201. https://doi.org/10.1016/j.diabres.2018.10.020

Mitchison, H.M., Valente, E.M., 2017. Motile and non-motile cilia in human pathology: from function to phenotypes. J. Pathol. 241, 294–309. https://doi.org/10.1002/

path.4843

Ow, G.S., Tang, Z., Kuznetsov, V.A., 2016. Big data and computational biology strategy for personalized prognosis. Oncotarget 7, 40200–40220. https://doi.org/10.18632/oncotarget.9571

Parimbelli, E., Marini, S., Sacchi, L., Bellazzi, R., 2018. Patient similarity for precision medicine: A systematic review. J. Biomed. Inform. 83, 87–96. https://doi.org/10.1016/j.jbi.2018.06.001

Peron, A., Vignoli, A., Briola, F.L., Morenghi, E., Tansini, L., Alfano, R.M., Bulfamante, G., Terraneo, S., Ghelma, F., Banderali, G., Viskochil, D.H., Carey, J.C., Canevini, M.P., TSC Study Group of the San Paolo Hospital of Milan, 2018. Deep phenotyping of patients with Tuberous Sclerosis Complex and no mutation identified in TSC1 and TSC2. Eur. J. Med. Genet. 61, 403–410. https://doi.org/10.1016/j.ejmg.2018.02.005

Powles-Glover, N., 2014. Cilia and ciliopathies: Classic examples linking phenotype and genotype—An overview. Reprod. Toxicol., 42nd Annual Conference of the European Teratology Society 48, 98–105. https://doi.org/10.1016/j.reprotox.2014.05.005

Radley, J.A., O'Sullivan, R.B.G., Turton, S.E., Cox, H., Vogt, J., Morton, J., Jones, E., Smithson, S., Lachlan, K., Rankin, J., Clayton-Smith, J., Willoughby, J., Elmslie, F.F., Sansbury, F.H., Cooper, N., Deciphering Developmental Disorders (DDD) Study, Balasubramanian, M., 2019. Deep phenotyping of 14 new patients with IQSEC2 variants, including monozygotic twins of discordant phenotype. Clin. Genet. 95, 496–506. https://doi.org/10.1111/cge.13507

Reiter, J.F., Leroux, M.R., 2017. Genes and molecular pathways underpinning ciliopathies. Nat. Rev. Mol. Cell Biol. 18, 533–547. https://doi.org/10.1038/nrm.2017.60
Robinson, P.N., 2012. Deep phenotyping for precision medicine. Hum. Mutat. 33, 777–780. https://doi.org/10.1002/humu.22080

Sang, L., Miller, J.J., Corbit, K.C., Giles, R.H., Brauer, M.J., Otto, E.A., Baye, L.M., Wen, X., Scales, S.J., Kwong, M., Huntzicker, E.G., Sfakianos, M.K., Sandoval, W., Bazan, J.F., Kulkarni, P., Garcia-Gonzalo, F.R., Seol, A.D., O'Toole, J.F., Held, S., Reutter, H.M., Lane, W.S., Rafiq, M.A., Noor, A., Ansar, M., Devi, A.R.R., Sheffield, V.C., Slusarski, D.C., Vincent, J.B., Doherty, D.A., Hildebrandt, F., Reiter, J.F., Jackson, P.K., 2011. Mapping the NPHP-JBTS-MKS protein network reveals ciliopathy disease genes and pathways. Cell 145, 513–528. https://doi.org/10.1016/j.cell.2011.04.019

Sharafoddini, A., Dubin, J.A., Lee, J., 2017. Patient Similarity in Prediction Models Based on Health Data: A Scoping Review. JMIR Med. Inform. 5, e7. https://doi.org/10.2196/medinform.6730

SHIRE, 2016. The Global Challenge of Rare Disease Diagnosis - The benefits of an improved diagnosis journey for patients.

Wang, F., 2015. Adaptive semi-supervised recursive tree partitioning: The ART towards large scale patient indexing in personalized healthcare. J. Biomed. Inform. 55, 41–54. https://doi.org/10.1016/j.jbi.2015.01.009

Weng, C., Shah, N., Hripcsak, G., 2018. Call for papers: Deep phenotyping for Precision Medicine. J. Biomed. Inform. 87, 66–67. https://doi.org/10.1016/j.jbi.2018.09.017
Williams, R., Kontopantelis, E., Buchan, I., Peek, N., 2017. Clinical code set engineering for reusing EHR data for research: A review. J. Biomed. Inform. 70, 1–13. https://doi.org/10.1016/j.jbi.2017.04.010

Wolf, M.T.F., Saunier, S., O'Toole, J.F., Wanner, N., Groshong, T., Attanasio, M., Salomon, R., Stallmach, T., Sayer, J.A., Waldherr, R., Griebel, M., Oh, J., Neuhaus, T.J., Josefiak, U., Antignac, C., Otto, E.A., Hildebrandt, F., 2007. Mutational analysis of the RPGRIP1L gene in patients with Joubert syndrome and nephronophthisis. Kidney Int. 72, 1520–1526. https://doi.org/10.1038/sj.ki.5002630

Yu, S., Liao, K.P., Shaw, S.Y., Gainer, V.S., Churchill, S.E., Szolovits, P., Murphy, S.N., Kohane, I.S., Cai, T., 2015. Toward high-throughput phenotyping: unbiased automated feature extraction and selection from knowledge sources. J. Am. Med. Inform. Assoc. JAMIA 22, 993–1000. https://doi.org/10.1093/jamia/ocv034

Zeng, Z., Deng, Y., Li, X., Naumann, T., Luo, Y., 2019. Natural Language Processing for EHR-Based Computational Phenotyping. IEEE/ACM Trans. Comput. Biol. Bioinform. 16, 139–153. https://doi.org/10.1109/TCBB.2018.2849968

Zhang, P., Wang, F., Hu, J., Sorrentino, R., 2014. Towards personalized medicine: leveraging patient similarity and drug similarity analytics. AMIA Jt. Summits Transl. Sci. Proc. AMIA Jt. Summits Transl. Sci. 2014, 132–136.

Zhang, X.A., Yates, A., Vasilevsky, N., Gourdine, J.P., Callahan, T.J., Carmody, L.C., Danis, D., Joachimiak, M.P., Ravanmehr, V., Pfaff, E.R., Champion, J., Robasky, K., Xu, H., Fecho, K., Walton, N.A., Zhu, R.L., Ramsdill, J., Mungall, C.J., Köhler, S., Haendel, M.A., McDonald, C.J., Vreeman, D.J., Peden, D.B., Bennett, T.D., Feinstein, J.A., Martin, B., Stefanski, A.L., Hunter, L.E., Chute, C.G., Robinson, P.N., 2019. Semantic integration of clinical laboratory tests from electronic health records for deep phenotyping and biomarker discovery. NPJ Digit. Med. 2. https://doi.org/10.1038/s41746-019-0110-4