Synthetic Patient Perspective Data for the Curation and Evaluation of Rare Disease Patient-Facing Technology

Emily Nielsen(✉), Tom Owen, Matthew Roach, and Alan Dix

Swansea University, Swansea, Wales
{e.e.nielsen,t.owen,m.j.roach,a.j.dix}@swansea.ac.uk

Abstract. Patient-facing technology to support rare disease patients seeking diagnosis has received comparatively little focus from the literature, despite the recognition of its importance. We hypothesise that this is due to the challenges presented when designing pre-diagnostic patient-facing technology within this area. A significant obstacle for research in this area is the lack of data which represents the patient's perspective. Existing data typically does not present the temporal aspects of diagnosis which are crucial to evaluate the diagnosis time of technology and consists of clinical terminology which is not representative of patients. This work aims to bridge this gap by creating open-source data which: (i) utilises patient-friendly terms and (ii) facilitates the sequencing of phenotypes to temporally recreate the informational journey of a rare disease patient. Therefore, this work facilitates evaluations on whether pre-diagnostic technology reduces the time to a rare disease diagnosis, thus providing more meaningful metrics for success.

Keywords: Rare disease · Patient-facing technology · Diagnosis · Health · Synthetic data · Data generation

1 Introduction

This paper looks at the generation of data for the evaluation of systems for rare disease diagnosis considering the need for (i) a temporal lens (importance of early diagnosis); and (ii) a patient-centred approach. Rare disease patients face a long and difficult journey to attain a diagnosis, resulting in severe, permanent and debilitating effects on their health [6]. This is often referred to as a diagnostic odyssey, with the average patient waiting four years, consulting with five clinicians, and receiving three misdiagnoses before they are correctly diagnosed [16]. Technologies to support rare disease diagnosis are almost always clinician-facing [7]. However, the role of the patient may be far more significant for rare conditions [4,6]. 94.6% of clinicians believe that they have insufficient or very poor knowledge of rare diseases [21] and lack time to research rare diseases. In contrast, it is common for patients to research their health, utilising resources such as ChatGPT, Google or Facebook [3,19]. It follows that patients may be able to contribute significantly to consultations with their healthcare providers. Indeed, The UK Strategy for Rare Diseases states that patients can play a significant role in diagnosis and treatment decisions *if given suitable resources* [4].

However, patients with rare diseases feel that they lack the support they need [5] which may be why patients resort to technology which is not specifically designed for health. In addition, applications designed with the common interest in mind will not cater for rare disease patients; information that is relevant to a rare condition is inevitably irrelevant for the majority. This suggests that patients do not have the resources required to play an active part in their health. Therefore, there exists a need for patient-facing pre-diagnostic technology which caters for the needs of people with undiagnosed rare diseases. Since this need has been established, we hypothesise that the limited focus in this area is due to the lack of data which is representative of the patient's perspective. Patient-facing works include Kühnle et al.'s [12] paper on the design of RarePairs, a peer-matched social media platform for rare disease patients. This required the use of an extensive 50-question survey to match patients based on their experiences through the healthcare system. The use of low-data approaches like this suggests that the lack of patient perspective data is a key barrier to data-driven approaches within this area.

Given that rare disease diagnosis is a long process where patients are unlikely to be diagnosed at the first consultation, a positive outcome for patients with rare conditions can be defined as the identification of a diagnosis as early in their journey as possible. As such, the performance of pre-diagnostic technology for rare disease patients needs to be evaluated throughout the diagnostic odyssey. Hence, test data must facilitate this temporal aspect of evaluation. However, as far as the authors are aware, there is only one paper [17] which evaluates the time taken for the proposed model to reach a correct diagnosis. Ronicke et al. evaluated their system, Ada DX, for each consultation that a patient has by manually removing data which would not be available for a given consultation. This revealed that only 33.3% of cases identified the correct condition in the top-5-fit disease list at the first consultation, however, Ada DX suggested the correct disease before clinical diagnosis for 53.8% of cases. This shows a non-trivial difference in the evaluation of this system since reducing the time to diagnosis for over half of the cases is highly significant in this context, however, to only show a single-point accuracy of 33.3% does not accurately portray the effectiveness of this system. Therefore, technology to support the diagnosis of rare diseases must be evaluated at multiple points to get an accurate impression of its effectiveness. While the evaluation approach presented by Ronicke et al. facilitates this, it may not be feasible in several research projects since the breakdown of cases into each of the clinical visits may not be possible (i.e., this information may not be present in the data, or it may be too time-consuming). Moreover, this approach aims to support evaluations of clinically-based technologies and thus is based on clinical data which does not represent the patient's perspective, so other methods may be required for patient-facing technologies. Therefore, there exists a need for approaches to evaluate pre-diagnostic technologies for rare diseases which can support the evaluation of patient-facing technology and does not require significant editing of data for each evaluation.

Hence, we identify two key barriers to accessing suitable data for the implementation and evaluation of rare disease patient-facing technology. Firstly, many sources of patient data are obtained from Electronic Health Records (EHR) [11,22] which consist of technical, clinical terminology. These data are not suitable for patient-facing technology as they do not use patient terminology for symptoms. Secondly, each of the numerous consultations involved in a rare diagnosis [6] presents an opportunity for diagnosis, so static performance metrics do not offer sufficient insight into the benefits of pre-diagnostic technology for rare diseases. Hence, a suitable test set must present different stages of the informational journey of each rare disease patient in order to assess the performance of pre-diagnostic technology over the stages of the diagnostic odyssey. Therefore, we need data which not only uses non-expert terminology, but that also sequences data in the order of a patient's information discovery. To address this gap, we present a data generation process to curate patient perspective data which we make freely available on GitHub[1]. This dataset provides a basis from which we generate: the Static User Profile Data which provides a static dataset for the curation of pre-diagnostic technology; and the Time-Series Persona Data which provides a suitable test set for evaluating pre-diagnostic technology for rare disease patients. These datasets aim to facilitate early-stage and proof-of-concept studies for rare disease pre-diagnostic technology.

2 Data Curation Process Overview

Figure 1 provides an overview of the processes to curate the datasets presented in this paper. The process begins with the established Orphanet rare diseases and phenotypes dataset, which contains phenotypes and the frequency of occurrence of each phenotype within each given disorder, and which is described in more detail later. We then augment Orphanet's data with patient perspective information (layman terminology and patient information discovery) to curate the Patient Perspective Dataset. This dataset provides a basis from which we generate the Time-Series Persona Data and the Static User Profile Data. We describe this process in more detail below.

2.1 Included Disorders

Creating a dataset for the estimated 10,000 types of rare diseases [13] presents a significant amount of work for proof of concept evaluations. Therefore, to establish a challenging but tractable basis on which to evaluate the potential of early-development pre-diagnostic technology, we adopt a subset of three

[1] https://github.com/902549/patient_perspective_data.

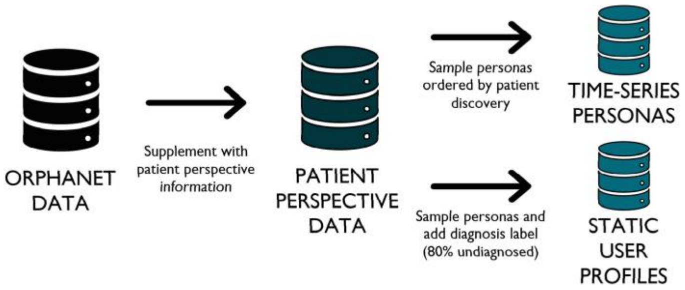
ORPHANET DATA
Supplement with patient perspective information
PATIENT PERSPECTIVE DATA
Sample personas ordered by patient discovery
TIME-SERIES PERSONAS
Sample personas and add diagnosis label (80% undiagnosed)
STATIC USER PROFILES

Fig. 1. Overview of the curation process for each of the three datasets

main rare diseases: Fabry Disease[2], Gaucher Disease[3], and hypermobile Ehlers-Danlos Syndrome (hEDS)[4] as the 'positive' classes for the test data. These were chosen because they are well documented, have varying diagnostic difficulties, and are easily misdiagnosed [2,9,15]. Thus, they provide a suitably challenging information-seeking task but also have sufficient documentation to scrape additional data. Clearly, for the purposes of evaluation, we need additional conditions in the databases to sufficiently assess their performance. The inclusion criteria were chosen to create a real-life distribution of the experimental context. That is, the three main rare diseases provided the basis for the Time-Series Persona Data (i.e., the positive classes), but prototypes should also include similar conditions (i.e., the negative classes - conditions that are most likely to be mistaken as the positive class).

The inclusion of misdiagnoses was guided by the relevant literature on the disease in question [1,10,14,15]. Psychological conditions were excluded and only specific named conditions were included. This resulted in a total of 16 rare disorders to provide the negative classes[5] (19 disorders including the three positive classes). Note that our data aims to facilitate comparative evaluations between completing systems, not aiming to give an absolute measure for any given system: therefore it is sufficient to have a range of conditions, but not match base-rate prevalence. In addition, our data aims to facilitate symptom-based patient matching systems which aim to cater for those facing a diagnostic odyssey, so people with common diseases would not have the need of this tool. The database

[2] https://rarediseases.org/rare-diseases/fabry-disease/ https://medlineplus.gov/genetics/condition/fabry-disease/ https://www.ninds.nih.gov/health-information/disorders/fabry-disease.

[3] wikipedia.org/wiki/Gaucher's_disease/ https://www.ncbi.nlm.nih.gov/books/NBK1269/.

[4] https://www.nhs.uk/conditions/ehlers-danlos-syndromes/ https://www.ncbi.nlm.nih.gov/books/NBK1279/ https://www.ehlers-danlos.com/what-is-eds/hypermobile-ehlers-danlos-syndrome-heds/.

[5] rheumatic fever, dermatomyositis, erythromelalgia, myelofibrosis, five rare forms of leukemia, two rare types of avascular necrosis, two rare forms of rheumatoid arthritis, vEDS, cEDS, cardiac-valvular EDS.

in a real-world context would be comprised of its users, so there would not typically be patients with common conditions in the database.

3 Curating the Patient Perspective Dataset

Both the Static User Profile Data and the Time-Series Persona Data require realistic patient data which represents the patient's perspective. However, many sources of patient data are obtained from Electronic Health Records (EHR) which consist of technical, clinical terminology [11,22]. These datasets show patients, their condition, and their phenotypic information, either stored as codes or in raw text. However, they are often difficult to access and do not represent a patient's perspective, since this would naturally consist of non-expert language. Some datasets exist which reflect the patient perspective, for example, a number of companies, such as Apple or Google, collect health data (e.g. symptom logging and sensor readings) from smart devices, but these datasets are considered proprietary information and as such are not publicly available. We therefore propose an approach to generate synthetic patient data consisting of non-expert terms for the phenotypes.

We need to create unique and realistic profiles to ensure both our Time-Series Persona Data and Static User Profile Data are as close to real data as possible, thus some form of real patient data must be used as a basis. Since data reflecting the patient perspective is not easily accessible, we use clinical patient data as a basis to create patient perspective profiles. Therefore, several aspects of the clinical patient data will need to be edited to make it representative of the patient's perspective. To ensure this process is efficient, we utilise a knowledge base to create the Patient Perspective Dataset (as opposed to data consisting of individual patient cases) to act as a base from which we can generate as many patient profiles we choose without creating additional work.

Orphanet has endeavoured to gather and improve knowledge on rare diseases since 1997, and as such has created several knowledge bases which provide insight into many different aspects of rare diseases. One knowledge base that Orphanet curated is their rare diseases and phenotypes knowledge base[6]. This contains phenotypes as well as the corresponding frequency of occurrence of each phenotype within each given disorder. In addition, it consists of standardised clinical terminology, is widely used for rare disease research, is frequently updated, multilingual, open source, and spans thousands of rare disorders. Moreover, it enables reproducibility and ensures that, if used in future development, it remains up to date and this process can be expanded to curate larger and more comprehensive datasets. Therefore, Orphanet's knowledge base provides a suitable starting point to create the Patient Perspective Dataset which provides the base from which we generate both the Time-Series Persona Data and the Static User Profile Data. Below we first describe the process used to create the Patient Perspective Dataset, then we present the remaining processes for the Static User Profile Data as well as the Time-Series Persona Data.

[6] https://www.orphadata.com/phenotypes.

Now that we have chosen the clinical data to base our patient profiles on, we need to alter this data to represent the perspective of rare disease patients. To create the Patient Perspective Dataset, we enrich the Orphanet data by augmenting the phenotype information for each disease with: HPO categories, layman terms, phenotype discovery group (i.e., development traits, symptoms, exploratory clinical findings, specific clinical findings), and probability of occurrence. Then we define a phenotype sampling process to dynamically generate a range of varied patient profiles from the Patient Perspective Data, using the probability of occurrence and some perturbation noise, each patient profile samples a proportional amount of phenotypes.

3.1 Identifying Patient Terminology for Phenotypes

For each phenotype associated with the 19 disorders, the knowledge base used standardised clinical terminology, namely Human Phenotype Ontology (HPO) terms [18]. HPO not only provides a standardised list of clinical terminology, but it also has synonyms and definitions and classifies phenotypes into categories of the human body (i.e., organ systems and parts of the body). All of this information is accessible on their website[7].

Using HPO's synonyms and definitions, we created initial patient terms. In particular, where synonyms were deemed to be patient- or non-expert-friendly, these were chosen. If there were no synonyms or only expert synonyms, the definitions were checked for short phrases which could be considered synonymous terms in themselves. Phenotypes that did not have clear patient-friendly terms from HPO were researched and discussed among multiple groups of two to four non-experts (people who were not healthcare professionals) until a term was unanimously agreed upon.

Once all terms were finalised by the non-experts, an experienced healthcare professional was shown the original HPO terms as well as the non-expert terms that we created for the three main disorders. They then checked that the patient phenotypes matched the original HPO term. The non-expert terms were then updated according to the suggestions made by the healthcare professional.

3.2 Labelling Phenotypes with Their Discovery Group and HPO Category

Now that we have finalised our layman phenotype terms, let us consider the order of discovery of the different types of patient data and group them accordingly. First, we define and categorise phenotypes into four different discovery groups: developmental traits, symptoms, exploratory clinical findings, and specific clinical findings. Second, we identify pre-requisites for specific clinical findings from existing phenotypes. Finally, we add additional pre-requisite symptoms for phenotypes which can be considered conditions in themselves identified by a clinician. These stages provide the crucial sequencing information which we later use to create the Time-Series Persona Data.

[7] https://hpo.jax.org/.

Define and Label Phenotypes by Discovery Group. We can consider each phenotype to either be symptoms (i.e., patient observable) or clinical findings (i.e., not observable by the patient). Since patients would clearly identify phenotypes they can observe themselves before those which require a clinical investigation, these phenotypes should come first. Some symptoms, such as *feeding difficulties in infancy*, would be observable from birth or a very young age. Therefore, it is important to distinguish these phenotypes, so the first two discovery groups are symptoms and developmental traits.

In addition, we can consider clinical findings to have two main types: exploratory clinical findings and specific clinical findings. We define exploratory clinical findings to be traits that are likely to be identified from routine investigations (e.g., blood tests, standard physical examinations). As such, many of these findings will be identified during or shortly after a patient visits their primary care physician.

We define specific clinical findings as traits which are likely to only be identified from specific investigations. It would be unrealistic for specific investigations to be conducted without the presence of symptoms to prompt these investigations. For example, the probability of discovering an *abnormal myocardium morphology* (abnormal heart wall muscle) is increased with the presence of cardiovascular symptoms, such as chest pain. Therefore, for each specific finding, we must ensure that the pre-requisite symptoms that are needed to prompt the necessary investigations are also sampled along with the specific finding.

Identify Pre-requisites for Specific Clinical Findings. This stage utilises HPO categories to identify pre-requisites for specific findings, namely the HPO category of a specific finding denotes the HPO category to identify required pre-requisites. That is, (i) we scrape HPO categories for each phenotype within the dataset, (ii) when a specific finding is sampled, 1-2 symptoms with the same HPO category are sampled.

First, we need to establish which symptoms should be considered pre-requisites for a given specific finding. To do this, we augment each phenotype in our data with its HPO category (i.e., organ systems and other physiological categories) as gathered from HPO's website. Then, for each sampled specific clinical finding, we additionally sample pre-requisite symptoms with the same HPO category as the finding. We generate one to two pre-requisite symptoms for each specific clinical finding using a random number generator to add perturbation noise. For example, if *abnormal myocardium morphology* was sampled, cardiovascular symptoms such as *heart palpitations* would also be sampled.

Add Clinician Identified Pre-requisites. In addition, some phenotypes, such as anaemia, may be considered conditions in of themselves. As such, there would naturally be symptoms associated with them, however, this was not present in the data. Since a clinician will know these conditions well, by denoting the presence of the condition, the associated symptoms may be implied. However, a patient will not necessarily recognise these associated symptoms. Therefore, we also add key symptoms associated with conditions that are present in the dataset. The underlying symptoms of phenotypes of this nature were identified by a healthcare professional and as such these symptoms were added and categorised into a new discovery group, pre-requisites. This ensured that they were only sampled if they are a pre-requisite of a phenotype that has been sampled, and otherwise cannot be sampled.

3.3 Probabilistic Sampling of Phenotypes

To generate a realistic range of phenotypes for each patient profile for a given disease we transformed unstructured frequency values from Orphanet's ontology into a probability density. The Orphanet ontology consisted of categorical values stored as strings (namely, always present: 100 %, very frequent: 99%-80% frequent: 79%-30%, occasional: 29%-5%, rare: 4%-1%, excluded: 0%). For values in a range, we took the mid-point of the percentage values and converted these to numbers between 0 and 1.

To generate unique patient profiles, we take a sample of phenotypes for a given disorder by randomly sampling from their defined frequency distribution using Numpy's random choice function[8]. Since the phenotypes in each discovery group are generated separately, we need to normalise our frequencies for each discovery group (this is a requirement of the sampling function). Since we sample within each discovery group, we lose the distribution of each group, so we set the number of phenotypes sampled to be representative of the weight of the discovery group's frequency. In addition, the proportion of phenotypes for each discovery group will vary from patient to patient, so we add some perturbation noise to the number of phenotypes sampled within the discovery group.

Now, let us formalise this as an equation for clarity. Given a disorder there is a set D of phenotypes for that disorder and a collection $G_1 ... G_5$ of the discovery groups for the disorder that have the properties: (i) $G_k \subset D$ – a disorder's discovery group only includes phenotypes for that disorder, (ii) $D = \bigcup(G_k)$ – every phenotype for the disorder is in a discovery group, (iii) $G_k \cap G_l = \emptyset$ for $k \neq l$ – no phenotype in more than one discovery group. So, we define the weight, w, of the discovery group as $w = \frac{\sum_{i \in G_k} F_i}{\sum_{i \in D} F_i}$, where F_i is the frequency of phenotype i. Therefore, given the total number of desired phenotypes per patient profile T, the size of phenotypes sampled within a specific discovery group is $T * w$ with some perturbation noise and then rounded to the nearest whole number.

4 Static User Profile and Time-Series Persona Data

Two separate processes follow now that we have created our Patient Perspective Data and defined our process for sampling phenotypes. For the Static User Profile Data, a diagnosis status and a name are assigned to the patient profile to create a user. For the Time-Series Persona Data, the phenotypes are sequenced based on their discovery group and divided into informational stages. We describe these steps in more detail in the following sections.

[8] https://numpy.org/doc/stable/reference/random/generated/numpy.random.choice.html.

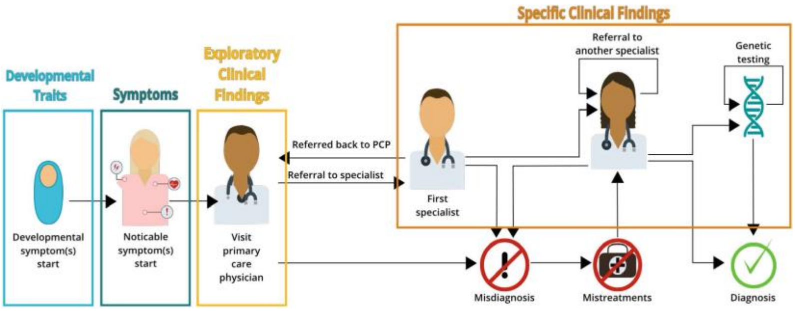
Developmental Traits
Symptoms
Exploratory Clinical Findings
Specific Clinical Findings
Developmental symptom(s) start
Noticable symptom(s) start
Visit primary care physician
Referred back to PCP
Referral to specialist
First specialist
Referral to another specialist
Genetic testing
Misdiagnosis
Mistreatments
Diagnosis

Fig. 2. Information discovery during a rare disease diagnostic odyssey

4.1 Creating the Static User Profile Data

Using the method described above, we sample patient personas from the Patient Perspective Dataset. The Static User Profile Data is intended to provide a user base for recommendation systems for symptom-based patient matching, so now we augment the data for this purpose. First, we set a diagnostic status for users. Let's consider a patient matching system that is predominantly aimed to be a pre-diagnostic application but not limited to pre-diagnosis. As such, we assume the majority of users are undiagnosed, so some people with diagnoses will be on the system. Therefore, we populate 80% of the users in the Static User Profile Data to be undiagnosed, the remaining users' diagnostic status displays the sampled disorder name. In addition, to ensure that the data presents humanistic users (as would be the case in a real-world context), we include randomly generated first names for each of the profiles. Therefore, the Static User Profile Data portrays a patient profile with a name, diagnosis status and a list of phenotypes.

4.2 Creating the Time-Series Persona Data

In this section, we outline our process to create the Time-Series Persona Data. The Patient Perspective Dataset that we generated above to provides a basis from which we can generate the Time-Series Persona Data. Given that the Patient Perspective Dataset was based on a rigorous knowledge base, it provides a realistic static basis for the Time-Series Persona Data. However, to simulate the informational journey of a rare disease patient, we need to augment the data with critical sequencing information which is representative of a patient's discovery of phenotypes. In particular, we use the sampling process described above to create patient personas. We then use the discovery groups to sequence phenotypes of each patient persona based on the patient information discovery of the phenotype group. These steps are described in more detail below.

First, we recreate the temporal aspect of information discovery from a patient's perspective. To do this, let us relate the phenotypes from our data to the diagnostic odyssey. In particular, as shown in Fig. 2, we can augment the patient journey diagram with the discovery groups, showing the order in which clinical discoveries are made and observed by the patient. The different types of data which we divided into the following discovery groups: developmental traits; symptoms; pre-requisite symptoms; exploratory clinical findings; specific clinical findings. This grouping can be utilised to provide the synthetic data with critical sequencing information to facilitate the revelation of phenotypes as each discovery group is perceived by the patient.

Patients discover the different types of phenotypes at different rounds of their diagnostic journey. Firstly, phenotypes which are observable by patients would be discovered first. Since developmental traits are present from an early age, it follows that these phenotypes should occur first, followed by non-developmental symptoms. Secondly, once the patient seeks medical help, clinicians will start providing them with information from tests or physical examinations. Routine investigations are often made on the first few visits, so exploratory findings will be identified first. Specific findings will occur latest in the diagnostic odyssey since these findings require specific investigations prompted by the phenotypes observed thus far. Therefore, to ensure that the Time-Series Persona Data is representative of a rare disease odyssey, we order the synthetic data so that developmental traits come first, followed by symptoms (including pre-requisites), exploratory findings, and finally specific findings.

Following this, we divided the phenotypes into different stages of information discovery by distributing the sorted phenotypes equally into the number of stages desired. In our case, we distributed the phenotypes equally to ensure that each stage was consistent in the amount of information that was revealed. Table 1 shows an example of the final generated Time-Series Persona Data for each of the three conditions included in the laboratory study.

Table 1. Time-Series Persona Data example for the three conditions

Condition	Early Stage	Middle Stage	Late Stage
hEDS	Sleep disturbance, Joint dislocation, Stretchy skin, Elbow dislocation, Muscle pain	Fatigue, Heartburn, Thin skin, Depressivity, Vertigo	Nausea and vomiting, Soft skin, Constipation, Gastrointestinal dysmotility, Extra bones in the cranium
Fabry Disease	Small dark-red spot, Poor appetite and weight loss, Lack of sweating, Nausea and vomiting	Joint pain, Blood in urine, Thickened skin, Vision loss	Kidney damage/Kidney disease, Cataract, Optic atrophy, Corneal dystrophy
Gaucher Disease	Squint, Corneal opacity, Joint pain, Abdominal pain, Recurrent fractures	Tremor, Falling, Delayed skeletal maturation, Recurrent fractures	Osteopenia: Decreased bone density, Enlarged liver, Cranial nerve paralysis, Enlarged spleen

5 Evaluating the Database

The steps above augmented a clinical knowledge base with patient terminology and sequencing information to create the Static User Profile Data and the

Time-Series Persona Data. These datasets provide data which represents the patient perspective for the curation of patient-facing technology. In addition, the Time-Series Persona data provides a temporal dataset to facilitate evaluations at different stages of the diagnostic odyssey. This provides more meaningful measures of the efficacy of technology by facilitating assessments on how quickly and consistently it suggests the correct diagnosis. However, before we use this data to perform evaluations, we must assess its suitability for purpose. In particular, we need to evaluate whether it provides a suitable representation of the informational journey of a real-world rare disease diagnosis.

Typical methods of evaluating synthetic data include comparing with other data using statistical methods or the performance of machine learning models. We do not have other data to compare our synthetic data to, so we could not use this evaluation method. Another method of validation is through feedback from domain experts, or by evaluating the utility of data for its intended application.

The Patient Perspective Dataset intends to represent patients, but is based on clinical data, so we can consider both healthcare professionals (HCPs) and non-HCPs to be domain experts in different ways. As non-HCPs would not know medical terminology, their input was important to provide terminology which was understandable to patients, whereas the HCPs knowledge of medical terminology verified that these terms were representative of the original HPO term. So, there was continual formative evaluation by domain experts throughout the process. As such, a subsequent expert evaluation of the Patient Perspective Data was not deemed necessary. Instead, we incorporate expert input to evaluate whether individual personas are realistic in the Time-Series Persona Data.

A healthcare practitioner with experience in primary care and rare diseases tested the Time-Series Persona Data by participating in a blinded simulation task, where the underlying condition of a given persona was only revealed at the end. They were presented three informational stages where new phenotypes were revealed and were asked to identify the condition using Google. Once the condition was revealed at the end of the three informational stages, they were asked about the realism of the task based on their clinical experience. They expressed that the patient persona made sense and was similar to their clinical experiences.

However, they discussed that when making diagnostic decisions, they would typically inquire about the patient's family history and duration of symptoms. Given that patients do not typically consider their family history until a genetic condition is suspected [8, 20], we did not consider this data feature to be necessary for the study. Moreover, we also deemed it could potentially lead participants to actively pursue genetic causes who would not have otherwise considered a genetic condition. The duration of symptoms, however, would be a relevant addition to the data, but as we did not have this information in the Orphanet dataset, we could not add this aspect. For future studies, we could explore methods to curate additional data features, such as the duration of symptoms. However, we deemed the current data to be sufficient for preliminary evaluations.

The Time-Series Persona Data may facilitate evaluations for a small number of conditions. In the context of proof of concept and early-stage evaluations, this provides a strong indication on potential. However, a larger dataset of patient phenotypes would be necessary to allow for evaluations on a greater scale. Large Language Models (LLMs) have shown significant promise in generating textual data based on a given input. Pre-trained transformer models, such as BART may be fine-tuned on the manually curated Patient Perspective Dataset, in addition to text scraped from HPO's website to provide a model which will translate clinical terminology to patient terminology. In addition, due to the variability of outputs from models like GPT 4, this may also provide multiple synonymous terms, thus adding to the realism of the patient data.

Future work may explore whether the Time-Series Patient Persona Dataset may be adapted to facilitate evaluations for clinician-facing technology. This may facilitate evaluation approaches that assess technology based on the time taken to diagnosis, rather than as a single-point accuracy. Ronicke et al. [17] performed a temporally-aware evaluation, however, it required significant manual edits to separate data based on the information that would be available at specific clinical visits. Researchers may not have sufficient time to perform these manual edits, so this dataset may provide a low-resource approach to facilitate temporal evaluations to support clinicians with rare diagnosis.

6 Conclusion

This paper presents a data generation approach for the curation of the Patient Perspective Dataset which aims to provide patient data which utilises (i) a temporal lens; and (ii) patient-centred language. We provide the Patient Perspective Dataset along with the code for sampling patient profiles on GitHub[9]. This data aims to bridge the gap where existing datasets were not suitable for patient-facing technology. Firstly, this data utilises non-expert terminology to represent the language used by patients. Secondly, this data is augmented with sequencing information to allow for temporal evaluations.

Clearly, the dataset presented in this paper provides a small sample of manually curated data. This provides a starting point from which a larger dataset could be curated with an automated pipeline. For example, a pre-trained transformer models may be fine-tuned on the Patient Perspective Dataset to provide a model which will translate clinical terminology to patient terminology. In addition, the HPO category, synonyms and definitions from HPO's website can easily be scraped using the phenotype's HPO ID and the disorder's Orphanet ID. Hence, this work can act as a basis from which an automatically generated patient perspective dataset may be curated. This would facilitate temporally aware evaluations to promote the design of pre-diagnostic technology which performs well throughout the stages of diagnosis. As such, this could open a new avenue of research to apply algorithms more suited to temporal contexts to this area which could prove more effective for this context of a lengthy diagnosis.

[9] https://github.com/902549/patient_perspective_data.

Considering temporal approaches for pre-diagnostic technology for rare diseases may increase the potential to support more challenging diagnoses where further investigations facilitate differentiation from common conditions.

Acknowledgments. The authors would like to thank Amicus Therapeutics for their support during this project. The main author is funded by the EPSRC Centre for Doctoral Training in Enhancing Human Interactions and Collaborations with Data and Intelligence Driven Systems (EP/S021892/1).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Berglund, B., Nordström, G., Lützén, K.: Living a restricted life with ehlers-danlos syndrome (eds). Int. J. Nurs. Stud. **37**(2), 111–118 (Apr 2000). https://doi.org/10.1016/S0020-7489(99)00067-X, https://www.sciencedirect.com/science/article/pii/S002074899900067X
2. Colomba, P., et al.: Fabry disease and multiple sclerosis misdiagnosis: the role of family history and neurological signs. Oncotarget **9**, 7758–7762 (2018). https://doi.org/10.18632/oncotarget.23970
3. De Choudhury, M., Morris, M.R., White, R.W.: Seeking and sharing health information online: comparing search engines and social media. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 1365–1376. CHI '14, Association for Computing Machinery, New York, NY, USA (Apr 2014). https://doi.org/10.1145/2556288.2557214
4. Department of Health UK: The UK strategy for rare diseases (2013). https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/260562/UK_Strategy_for_Rare_Diseases.pdf
5. Depping, M.K., Uhlenbusch, N., von Kodolitsch, Y., Klose, H.F.E., Mautner, V.F., Löwe, B.: Supportive care needs of patients with rare chronic diseases: multi-method, cross-sectional study. Orphanet J. Rare Dis. **16**, 44 (2021). https://doi.org/10.1186/s13023-020-01660-w
6. Faurisson, F.: Survey of the delay in diagnosis for 8 rare diseases in Europe: Eurordiscare2 (2004). https://www.eurordis.org/wp-content/uploads/2009/12/EURORDISCARE_FULLBOOKr.pdf
7. Faviez, C., et al.: Diagnosis support systems for rare diseases: a scoping review. Orphanet J. Rare Dis. **15**(1), 94 (2020). https://doi.org/10.1186/s13023-020-01374-z
8. Genetic Alliance, The New York Mid-Atlantic Consortium for Genetic and Newborn Screening Services: Understanding Genetics: A New York, Mid-Atlantic Guide for Patients and Health Professionals. Genetic Alliance, Washington (DC) (Jul 2009). https://pubmed.ncbi.nlm.nih.gov/23304754/
9. Halverson, C.M.E., Cao, S., Perkins, S.M., Francomano, C.A.: Comorbidity, misdiagnoses, and the diagnostic odyssey in patients with hypermobile ehlers-danlos syndrome. Genetics Med. Open **1**(1), 100812 (Apr 2023). https://doi.org/10.1016/j.gimo.2023.100812, https://www.sciencedirect.com/science/article/pii/S294977442300821X

10. Hershenfeld, S.A., et al.: Psychiatric disorders in Ehlers-Danlos syndrome are frequent, diverse and strongly associated with pain. Rheumatol. Int. **36**(3), 341–348 (2016). https://doi.org/10.1007/s00296-015-3375-1
11. Kruse, C.S., Smith, B., Vanderlinden, H., Nealand, A.: Security techniques for the electronic health records. J. Med. Syst. **41**, 127 (2017). https://doi.org/10.1007/s10916-017-0778-4
12. Kühnle, L., Mücke, U., Lechner, W.M., Klawonn, F., Grigull, L.: Development of a social network for people without a diagnosis (rarepairs): Evaluation study. J. Med. Internet Res. **22**(9), e21849 (Sep 2020). https://doi.org/10.2196/21849, http://www.jmir.org/2020/9/e21849/
13. Haendel, M., et al.: How many rare diseases are there? Nat. Rev. Drug Discov. **19**(2), 77–78 (2020). https://doi.org/10.1038/d41573-019-00180-y, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7771654/
14. Mehta, A., et al.: Fabry disease defined: baseline clinical manifestations of 366 patients in the fabry outcome survey. Europ. J. Clin. Invest. **34**(3), 236–242 (2004). https://doi.org/10.1111/j.1365-2362.2004.01309.x
15. Mistry, P.K., et al.: A reappraisal of Gaucher disease-diagnosis and disease management algorithms. Am. J. Hematol. **86**(1), 110–115 (2011). https://doi.org/10.1002/ajh.21888
16. Muir, E.: The rare reality - an insight into the patient and family experience of rare disease (2016). https://www.raredisease.org.uk/media/1588/the-rare-reality-an-insight-into-the-patient-and-family-experience-of-rare-disease.pdf
17. Ronicke, S., Hirsch, M.C., Türk, E., Larionov, K., Tientcheu, D., Wagner, A.D.: Can a decision support system accelerate rare disease diagnosis? Evaluating the potential impact of Ada DX in a retrospective study. Orphanet J. Rare Dis. **14**(1), 69 (2019). https://doi.org/10.1186/s13023-019-1040-6
18. Köhler, S., et al.: The human phenotype ontology in 2021. Nucleic Acids Res. **49**(D1), D1207–D1217 (2021). https://doi.org/10.1093/nar/gkaa1043
19. Shahsavar, Y., Choudhury, A.: User intentions to use chatgpt for self-diagnosis and health-related purposes: Cross-sectional survey study. JMIR Hum Factors **10**, e47564 (May 2023). https://doi.org/10.2196/47564, http://www.ncbi.nlm.nih.gov/pubmed/37195756
20. Walker, H.K., Hall, W.D., Hurst, J.W.: Clinical Methods: The History, Physical, and Laboratory Examinations. Butterworth-Heinemann Ltd, Boston, 3rd edn. (Apr 1990), https://www.ncbi.nlm.nih.gov/books/NBK201/, chapter 215 The Family History
21. Walkowiak, D., Domaradzki, J.: Are rare diseases overlooked by medical education? awareness of rare diseases among physicians in Poland: an explanatory study. Orphanet J. Rare Dis. **16**, 400 (2021). https://doi.org/10.1186/s13023-021-02023-9
22. Wen, Q., Ouyang, Z., Zhang, J., Qian, Y., Ye, Y., Zhang, C.: Disentangled dynamic heterogeneous graph learning for opioid overdose prediction. In: Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. p. 2009-2019. KDD '22, Association for Computing Machinery, New York, NY, USA (2022). https://doi.org/10.1145/3534678.3539279