PhenoLines: Phenotype Comparison Visualizations for Disease Subtyping via Topic Models

Michael Glueck, Mahdi Pakdaman Naeini, Finale Doshi-Velez, Fanny Chevalier,
Azam Khan, Daniel Wigdor, Michael Brudno

Fig. 1. PhenoLines facilitates the visual analysis of topics that describe disease symptoms, in support of topic model optimization and characterization. Hierarchical relationships, temporal trends, correlated measures, and rank-ordered lists enable for comparisons within and between topics. The interface includes (A) Settings Panel, (B) Detail Panel, (C) Topics Panel, and (D) Search Panel.

Abstract—PhenoLines is a visual analysis tool for the interpretation of disease subtypes, derived from the application of topic models to clinical data. Topic models enable one to mine cross-sectional patient comorbidity data (e.g., electronic health records) and construct disease subtypes—each with its own temporally evolving prevalence and co-occurrence of phenotypes—without requiring aligned longitudinal phenotype data for all patients. However, the dimensionality of topic models makes interpretation challenging, and de facto analyses provide little intuition regarding phenotype relevance or phenotype interrelationships. PhenoLines enables one to compare phenotype prevalence within and across disease subtype topics, thus supporting subtype characterization, a task that involves identifying a proposed subtype's dominant phenotypes, ages of effect, and clinical validity. We contribute a data transformation workflow that employs the Human Phenotype Ontology to hierarchically organize phenotypes and aggregate the evolving probabilities produced by topic models. We introduce a novel measure of phenotype relevance that can be used to simplify the resulting topology. The design of PhenoLines was motivated by formative interviews with machine learning and clinical experts. We describe the collaborative design process, distill high-level tasks, and report on initial evaluations with machine learning experts and a medical domain expert. These results suggest that PhenoLines demonstrates promising approaches to support the characterization and optimization of topic models.

Index Terms—Developmental disorder, Human Phenotype Ontology (HPO), Phenotypes, Topic models, Topology simplification.

1 Introduction

The characterization of complex developmental disorders, such as autism spectrum disorder (ASD), is a challenging and important task. In such diseases, the heterogeneity in symptom presentation and a lack of definitive diagnostic tests cause difficulties for diagnosis and prognosis: clinicians must rely on their experience to identify the disorder by observing the co-presentation of symptoms (i.e., *symptom comorbidity*), as well as make predictions about how the disease will evolve for a particular patient to determine treatment and care.

- *M. Glueck is with Autodesk Research and University of Toronto (E-mail: mglueck@dgp.toronto.edu)*
- *M. Pakdaman Naeini and F. Doshi-Velez are with Harvard University (E-mail: pakdaman@g.harvard.edu; finale@seas.harvard.edu)*
- *F. Chevalier is with Inria (E-mail: fanny.chevalier@inria.fr)*
- *A. Khan is with Autodesk Research (E-mail: azam.khan@autodesk.com)*
- *D. Wigdor is with University of Toronto (E-mail: daniel@dgp.toronto.edu)*
- *M. Brudno is with University of Toronto and Hospital for Sick Children (E-mail: brudno@cs.toronto.edu)*

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.201x.xxxxxxx

Symptoms are described by *phenotypes*—observable and measurable traits of patient morphology (e.g., enlarged heart), physiology (e.g., seizures), or behavior (e.g., depression)—and diseases are characterized by the probability of phenotype co-presentations. Characterizing diseases is challenging for complex developmental disorders because multiple disease processes may result in similar symptoms, and a single disease process may present differently in each child. For example, two children may have similar core ASD symptoms (e.g., language disorders), but only one may have severe gastrointestinal symptoms. Do the patients belong to different disease subtypes, or does one have two unrelated disorders? Symptoms may also vary as a child ages. An underlying neurological condition may produce convulsions in infancy and intellectual disability in childhood. Developing a robust understanding of how phenotypes manifest and evolve over time for a particular disease (i.e., *disease natural history*) is critical to improving the accuracy of diagnosis and prognosis.

To support the study of disease natural history, researchers have applied machine learning approaches to characterize diseases based on electronic health record (EHR) data via clustering [16], deep learning [8], support vector machines [35], and topic models [20]. In contrast to traditional longitudinal cohort studies of several hundred patients, machine learning enables analyses of tens of thousands of patients, thereby improving the differentiation of correlations between

heterogeneous phenotypes from spurious co-presentations of symptoms. This work focuses on improving the analysis workflow for topic model-based approaches to disease characterization.

Topic models can be trained on the symptoms of patient subsets across many ages, and then linked to create ad hoc disease history models based largely on cross-sectional data (i.e., EHRs). The resulting topics estimate the temporal evolution of phenotype probabilities that describe latent disease processes. Investigating the pattern of dominant phenotypes across topics can yield characterizations of disease subtypes [17, 20]. Consulting medical experts, who contribute domain knowledge to characterize and evaluate the clinical validity of topic models, can inform directions for debugging and refining the topic models. During this iterative cycle of model characterization and optimization, it is critical to investigate the patterns of modeled phenotypes both **within**, and **between** disease subtypes—e.g., *what phenotypes are most prevalent in each disease subtype*, *what phenotypes are common in some subtypes but not in others*, and *how do the prevalence of phenotypes in a disease subtype change over time*.

Due to the high dimensionality of topic models, de facto analyses often summarize topics using a subset of the modeled phenotypes (e.g., highest probability). These topic summaries may not accurately represent the semantic relationships between phenotypes. Phenotypes are not isolated, rather, they are indicative of higher-level biological systems (e.g., general neurological abnormalities). Each modeled phenotype also reflects a temporal evolution of probabilities (e.g., trends) that should be evaluated in tandem with the patterns of phenotypes in topics. Existing topic model analyses do not represent modeled phenotypes using hierarchical semantic relationships, and no tools exist to visualize the hierarchical and temporal dimensions of phenotypes through tightly-coupled interactive exploration.

We address the representation of semantic relationships by leveraging the taxonomy of the Human Phenotype Ontology (HPO) [32], which represents an anatomical hierarchy of subclass relationships between phenotypes, including multiple-inheritance (i.e., one or more superclass phenotypes). Using these relationships, the probabilities of modeled phenotypes can be aggregated to more general phenotypes, revealing higher-level patterns localized in biological systems. This enables disease subtypes to be characterized by considering phenotypes at multiple granularities (i.e., specific and general).

To this end, we contribute a data transformation workflow that processes unstructured topic model output into a graph-based abstraction using the topology of the HPO. This graph unifies the representation of both hierarchical and temporal aspects of modeled phenotypes. Yet, this graph poses challenges for visualization, as topic model output comprises thousands of modeled phenotypes, each with a temporal component, and a complex hierarchical structure. A novel measure of relevance was developed to simplify the graph topology through filtering and compression, by leveraging existing prevalence data annotated to the HPO. We demonstrate the graph can be transformed into visualizations for different vectors of analysis (i.e., radial hierarchy, timeline charts, scatterplots, summary table, and rank-ordered list).

In this work, we present PhenoLines, a visual analysis tool for the interactive interpretation of disease subtypes derived via topic models, based on this data transformation workflow. PhenoLines was designed to support machine learning experts when optimizing topic models, and facilitate characterization sessions between machine learning and medical experts. We describe the collaborative design process that we undertook to develop PhenoLines. Finally, we report the results of an initial evaluation using topic models derived from a dataset of ASD [16]. The evaluations include use by a collaborator, feedback from two third-party machine learning experts, and an in situ characterization session with a development-behavioral pediatrician. PhenoLines is an open source project; visit www.phenolines.org for an online demonstration, link to the GitHub repository, and license information.

2 Background and Related Work

Our work builds upon research in machine learning and the visualization of topic models and longitudinal medical data.

2.1 Topic Models of Clinical Data

Topic modeling is a machine learning approach for modeling discrete admixtures, first popularized for extracting themes from document collections [3]. The underlying assumption is that a dataset can be modeled as a set of *topics*[1], each of which are probability distributions over words—or in our case, phenotype terms. For example, a topic about neurological disorders may have high probability on phenotype terms such as seizures or migraines. The presentation of phenotypes for any patient can then be represented as a mixture of topics. For example, a patient with only neurological symptoms might have a high weighting associated to a topic about neurological disorders, while a patient with neurological and gastrointestinal concerns may be best represented by a mixture of two or more topics. The weighting applied to each topic can help label and cluster patients with similar disease presentations.

Topic models have been extended to describe the temporal evolution of topics [2, 54]. When applied to disease characterization, these topic models capture the evolution of phenotype prevalence as a patient ages. Thus, the resulting topics can help identify insightful presentations of a disease's progression or to differentiate disease subtypes. Topic models have been shown to be effective for characterizing diseases [17] and predicting patient outcomes in intensive care units [28, 42].

Visualization can address the challenge of interpreting and validating topic models [7]. In our work, we address interpretation and validation through visualizations that facilitate characterization and inform optimization. The iterative cycle of characterization and optimization identifies phenotype patterns that are inconsistent with the expectations of medical experts, enabling machine learning experts to target improvements to the topic model.

2.2 Visualizing Topic Models

The popularity of topic model-based text summarization motivated the design of a number of visualization systems to interpret document collections. Topic models have been leveraged in visualizations to organize documents into meaningful dimensional projections [9, 29, 41], clusters [34], and relationship networks [25]. Topic modeling has also been tightly integrated with interactive visualization to describe document collections (e.g. TIARA [37], ParallelTopics [18], TextFlow [13]). These systems assume a flat organization of topics, and use a river-flow metaphor or Sankey diagrams to visualize the temporal evolution of topics over time. As the number of topics increases, these representations become visually cluttered. To address scalability, hierarchical organizations of the evolution of topics were proposed (e.g. HierarchicalTopics [19], RoseRiver [14], TopicPanorama [36]). These approaches rely on the automated or semi-automated construction of topic hierarchies by clustering topics based on word similarity.

Disease subtype analysis can be distinguished from the thematic analysis of documents based on the focus on words, rather than topics. Themes summarize how **topics** evolve in relation to one other (e.g., themes relating to war or recessions wax and wane corresponding to historical events). In contrast, disease subtypes characterize the evolution of **words** (i.e., phenotype terms) across the time steps of individual topics. While prior work uses hierarchy to organize clusters of related topics, our approach uses a hierarchy to classify the phenotype terms within topics. A hierarchy produced by clustering is based on the words in the topics, and optimizations to the topic model may yield different hierarchies as the topics change. In contrast, the taxonomy of the HPO is a hierarchy that is independent of the topic modeling process, and thus consistent for different topic models, while the subclass relationships represent a classification of phenotype terms based on domain consensus. In summary, the hierarchy of the taxonomy not only summarizes probabilities to more general phenotype terms, but also communicates how phenotype terms relate to each other across different biological systems.

Topic models do not automatically provide meaning—they must be manually interpreted and evaluated by domain experts [7]. The Topic Browser [22], Termite [12], LDAvis [50] and LDAExplore [21] focused on verifying model quality through visual comparisons of how well

[1] We will use *topic*, *disease process*, and *disease subtype* interchangeably.

topics relate to each other and how well terms associate with each topic. Our work most closely relates to visual analysis tools that support the manual inspection and verification of the relevance and meaningfulness of latent topics. Chuang et al. [11] proposed a framework to support the large-scale assessment of topic relevance by aligning a set of latent topics and a set of reference topics, visualized as a correspondence chart. Since no reference topics exist in our domain, we propose a novel approach to estimate phenotype relevance by leveraging existing prevalence estimates via the HPO.

2.3 Visualizing Longitudinal Medical Data

A wealth of visualization tools has been developed to explore patient EHRs. Most efforts focus on patient medical histories in a time-oriented context, including monitoring a patient's condition (e.g., LifeLines [43], VisuExplore [45]), analyzing response to a treatment (e.g., IPBC [10], CareCruiser [26]), and comparing evolving symptoms to a baseline (e.g., Lifelines2 [53], LifeFlow [56]). See Rind et al. [46] and Shneiderman et al. [49] for reviews in this area. Clinicians find these tools helpful when determining a course of treatment for acute or chronic patient problems. Visualizations have also been developed to investigate data from longitudinal cohort studies, such as for the iterative refinement of event queries (e.g., CAVA [57]), to define temporal cohort membership constraints (e.g., COQUITO [33]), or to evaluate the effectiveness of treatments (e.g., CoCo [38]). Unlike the present work, these tools focused on extracting and visualizing sequences of events about specific patient events from EHR data. Our work focuses on the comparing and interpreting the temporal evolution of phenotype probabilities within and between disease subtypes. To our knowledge, our work is among the first visualizations of topic models applied to disease characterization.

3 Collaborative Design Process

The design of PhenoLines was a collaboration between researchers in visualization and machine learning. The collaboration was initiated because the machine learning researchers expressed an interest in applying the taxonomy of the HPO to help visualize their topic models. In this section, *we* is used to describe the design process from the perspective of the visualization researchers. To facilitate a systematic design process, we adapted the nine-stage design study methodology framework [48]. PhenoLines was developed over six months, during which we worked closely with the machine learning researchers, who took on the role of domain experts. Here, we summarize the four stages of the *Core Phase* of the design process.

Problem Characterization. In the *Discover Stage*, we investigated the problem of disease subtyping via topic models. First, we conducted two formative interviews with the machine learning researchers to learn about their research, observe existing tools, and discuss barriers to analysis. We followed-up with two semi-structured interviews to elicit specific details about their data, their workflow, and the types of analysis goals that could be supported through visualization. The transcripts of these interviews were coded to identify needs and requirements, which were distilled into abstract visualization tasks (Section 4). Two core requirements emerged: the hierarchical visualization of disease subtypes to identify interesting phenotypes, and the need to compare the temporal evolution of phenotype probabilities between subtypes.

Data Abstraction and Visual Encoding. In the *Design Stage*, we worked closely with the machine learning researchers to develop a graph-based data abstraction based on the taxonomy of the HPO, and evaluate visual encoding strategies. The machine learning researchers provided us with output from a topic model they had previously analyzed. This enabled iterative development and refinement of the data abstraction and visual encodings, as the machine learning researchers could confirm when expected patterns and insights were revealed.

The data abstraction addressed the requirement of hierarchical representation. The modeled phenotypes of each disease subtype became leaf nodes, and the probabilities were aggregated to more general phenotypes in the hierarchy. Working with the machine learning experts, we tested different measures to summarize the temporal evolution of phenotype probabilities as a single value that could be explicitly encoded in visualizations. Based on feedback from the machine learning experts, the measures of maximum probability, trend slope, and representative age interval (peak probability) best helped identify potential phenotypes of interest.

Implicit hierarchies (i.e., space-filling) were preferred over explicit representations because they resembled heatmaps; compactly encoding the node values and revealing patterns in the hierarchy (e.g., branches with similar values). Radial hierarchies were selected because they more evenly allocated size across all levels of the hierarchy, which was important because leaf nodes were crucial to the analyses.

To compare the temporal evolution of phenotype probabilities between subtypes, we tested explicit encodings of the differences (e.g., variance, mean-squared error, entropy). The machine learning experts found these measures were inflexible for the detailed comparisons necessary for analysis. They preferred juxtaposed timeline charts to compare the probabilities of phenotypes because the output from the topic model was directly represented. These timeline charts supported flexible comparisons as the need arose, making it easier to pinpoint specific differences that could be hidden in a summary measure.

Through the collaborative design work, it became clear that hierarchical and temporal representations alone did not provide all the necessary perspectives on the data. To mimic the existing workflow of the machine learning researchers, we incorporated a rank-ordered list of the phenotypes in each topic model to support summarization. As we developed the summary measures, our collaborators expressed a desire to visualize correlations between the measures within and between topics, resulting in the scatterplot and summary chart. The data transformation workflow and the computation of summary measures are described in Section 5.

Initial Prototype and Refinement. In the *Implement Stage*, we developed an interactive prototype using D3 [5]. The machine learning researchers were asked to evaluate features and report on bugs. During this stage, the Settings Panel (Fig. 1A) was developed to address configuration of the visualizations. The relevance score and hierarchy filtering and compression methods were developed to address visual complexity. We continued to refine and extend the prototype over four weeks to address usability issues. With each update, we were able to attain richer feedback due to improved features and usability. Iterative refinements continued until the machine learning researchers were satisfied they could use the prototype to make sense of, hypothesize about, and investigate the models they had produced. The resulting functional prototype is described in Section 6.

Functional Prototype Deployment. In the *Deploy Stage*, the functional prototype was evaluated by the machine learning researchers and two third-party machine learning experts. The prototype was also used by the machine learning researchers to conduct a disease characterization session with a medical expert. These evaluations address the top three levels of Munzner's Nested Model for Design and Validation [40]. Details of these evaluations are found in Section 7.

4 Disease Subtype Analysis

The problem characterization yielded insights into disease subtyping via topic models, key tasks, and existing analysis methods and challenges. We synthesize how visualization could improve the analysis workflow, and characterize the domain tasks as abstract visualization tasks.

4.1 Clinical Data and Topic Models

The machine learning researchers expressed a desire to characterize and optimize new topic models generated from an existing dataset of autism spectrum disorder (ASD) patients collected from Boston Children's hospital [16]. This dataset included the comorbidity information of 13,337 patients extracted from a corpus of 66,275 EHRs, and was preprocessed to represent the comorbidity data using unique term identifiers (CUI codes) from the Unified Medical Language System (UMLS) [4]. The UMLS is a meta-thesaurus that provides terminological mappings between a variety of medical vocabulary systems, including ICD, SNOMED CT, and HPO.

The machine learning experts explained the comorbidity data was partitioned into 15 mutually exclusive age time steps (0 to 14 years)

for the purposes of topic modeling. The CUI codes associated with each patient at each time step was considered a document using a bag-of-words representation. To reduce noise in the model, only the 5,000 most frequent CUI codes in the corpus were retained. A standard Latent Dirichlet Allocation (LDA) [3] model was applied to each of the 15 time steps independently. For each time step, a seven-topic LDA was trained using gensim [44], with both symmetric priors and hyper-parameter optimization. This process resulted in seven topics at each time step. The Kuhn-Munkres algorithm [39] was used to create a bipartite matching of topics at adjacent time steps, resulting in seven disease process models, each describing the phenotype probabilities for a disease subtype over the 15 time steps. The machine learning researchers provided us with the final output of the topic model, in the form of tuples (Topic ID, Time ID, CUI, and Probability).

4.2 Key Tasks

The formative interviews revealed that topic model researchers engage in an iterative process of topic model characterization and optimization.

4.2.1 Disease Subtype Characterization and Validation

The interviewees explained that the primary goal when interpreting their topic models is the characterization of modeled disease subtypes. This process involves an evaluation of the dominant phenotypes in each of the topics, but also benefits from a higher-level understanding of which biological systems are affected—e.g., dominant neurological phenotypes may indicate a broad correlation between these symptoms, or point to specific developmental disorders or psychiatric conditions.

During this task, machine learning researchers often consult medical domain experts to ascertain whether the characterizations derived from the topic models align to known comorbidities of the disease, and whether they reveal new correlations that were not expected. These conversations can also shed light on whether novel patterns are scientifically interesting, or reveal biases present in the dataset.

4.2.2 Topic Model Debugging and Refinement

Topic models are optimized by improving prior estimates of phenotype prevalence. The interviewees explained that due to the non-linearity of the search space in topic modeling, the model results must be checked to ensure unexpected effects are not introduced whenever the topic model is modified. The output also needs to be verified to ensure that existing features continue to be captured. If the refined model includes unexpected changes in the definition of topics, it is necessary to drill down into the details to investigate where the changes originate.

Feedback from the topic characterization efforts can also guide optimizations. For example, one interviewee explained that a common issue transpires when a phenotype occurs with relatively high probability across multiple topics—in some cases, this pattern may be perfectly valid, indicating that the phenotype is common but irrelevant to differentiating disease subtypes. In other cases, it may indicate that a poor choice of prior estimates prevented differentiation.

4.3 Existing Analysis Methods and Barriers

The machine learning researchers explained that specialized tools for in-depth exploration of disease topic models are currently unavailable. They typically write custom analysis scripts to summarize topics based on the modeled phenotypes with highest probability in each topic and at each time step. Due to the large number of phenotypes, only a fixed number of the highest probability phenotypes are typically evaluated. They noted a shortcoming of this approach is that it does not consider higher-level affected biological systems. The interviewees reported they manually produce timeline charts of temporally evolving probabilities to investigate the trends of specific phenotypes, but this is a tedious, error-prone, and time-consuming process.

This largely manual process of analysis impedes fluidity of both model characterization and optimization tasks. The interviewees explained that manually creating timeline charts for subsets of phenotypes impacted the quality of disease subtype characterization sessions with medical experts. Since charts could not be prepared for all phenotypes, if discussions with medical experts required details of an unexpected phenotype, follow-up sessions were necessary to complete the characterization, which broke continuity of the discussion. They also noted that the manual analyses made it difficult to maintain an overview of phenotype probabilities across disease subtypes, impacting their ability to track the full effects of topic model optimizations.

4.4 Benefits of Interactive Visualization

The de-facto approach of displaying only the highest probability phenotypes, without consideration of their hierarchical relationships, can result in incorrect scientific interpretation. For example, there are over 20 different phenotypes for convulsion-related phenotypes. A patient with epilepsy will accrue combinations of these 20 phenotypes depending both on the details of their sub-condition as well as their clinician. Thus, a topic relating to patients with epilepsy may put 0.01-0.03 probability on each individual term, which, when summed result in 0.2-0.3 of the topic having to do with epilepsy (with the rest of the probability mass having to do with comorbid conditions). In contrast, terms like intellectual disability and autism have many fewer associated phenotypes. Thus, a topic relating to patients with these conditions may put much higher probabilities—0.1-0.2—on individual phenotypes. If one looked only at the highest probability phenotypes in such a topic, we might mistakenly believe that the topic was only about autism and intellectual disability because all other phenotypes had probability less than 0.05. The hierarchical representation of phenotypes in topics using the taxonomy of the HPO can provide context that reveals relationships between individual phenotypes and can also help identify whether topics differentiate particular biological systems. This approach could improve on the existing summarization approaches used by the machine learning researchers.

An interactive visualization system can also address the challenge of fluidity in characterization and optimization tasks by providing a consistent visual representation of the topics, making it easier to track changes resulting from optimizations. Characterization sessions can benefit through an interface to view details on demand for any phenotype in the topic model.

4.5 Visualization Tasks

We classified the identified domain tasks in terms of abstract visualization tasks using Brehmer & Munzner's Multi-Level Task Typology [6].

Considering *Why?* users perform tasks, both aim to **discover**, but differ in *search* and *query* approaches to interpret the underlying data. The Characterization and Validation task is oriented toward identifying specific phenotype co-presentations and trends in the temporal probability distribution. In this case, the target is known, so *search* involved **locate** (location known) and **explore** (location unknown). Through *query*, the goal of this task is to **identify** and **compare** the phenotypes within topics. In contrast, the Refinement and Debugging task is primarily concerned with the high-level patterns of phenotype temporal probability distributions between topics. In this case, the target is unknown, so *search* involves **browse** (location known) and **explore** (location unknown). Through *query*, the goal of this task is to **compare** or **summarize** the differences between topics.

To address *How?* users complete tasks, our design process revealed the need to **encode** unstructured topic models output in hierarchical and temporal visual representations, among others. We thus developed a graph-based data abstraction that can be transformed into task oriented visual representations. Using this graph, summary measures of the temporal evolution of phenotype probabilities were developed to compare and summarize broader trends between topics, and also identify specific phenotypes within topics. Visual encoding strategies were used to represent specific details to support comparisons. For example, the summary table and timeline charts enable comparison through juxtaposition, while scatterplots and rank-ordered lists explicitly represent relationships between data attributes. Due to the variety of visual representations, users **manipulate** the visualizations through *select*, *arrange*, and *change* operations that maintain a consistent layout, rather than navigation. To address visual clutter and complexity, we support *filter* and *aggregate* operations by developing a novel measure of phenotype relevance.

5 Data Transformation Workflow

The interactive analysis of disease topic models required a data abstraction that represented the hierarchical and temporal aspects of the topics. To this end, we developed a workflow to map CUI codes from the topic model output to phenotype terms in the HPO, and leverage the taxonomy of the HPO to derive a graph-based data abstraction of the modeled phenotypes.

5.1 Human Phenotype Ontology (HPO)

The HPO is one of the most broadly used ontologies for phenotypes. As an on-going initiative, the HPO standardizes terminology and defines relationships between phenotypes (i.e., semantic, logical, hierarchical). The HPO facilitates interoperability with external resources that link genes, phenotypes, and diseases (e.g., OMIM, Orphanet) [47] and currently includes over 11,000 terms with over 250,000 annotations to rare and common diseases. Although other medical nomenclatures (e.g., ICD, SNOMED CT) have hierarchical structures, those categorizations are oriented around medical billing or exhaustive lists of medical terms, including procedures and anatomical parts. In contrast, the HPO classifies the semantic relationships between phenotypes, directly representing symptoms of diseases within biological systems.

Leveraging the structure and external resource integration of the HPO enables computation on phenotypes that is not possible using current EHR coding nomenclatures alone (e.g., ICD, SNOMED CT). For example, similarity scores between patients with non-overlapping phenotypes can be calculated and concepts like the diagnostic significance of a phenotype can be quantified [31]. In prior work, we demonstrated the application of the HPO to visualize phenotype data within a hierarchical semantic context [23, 24]. In this work we extend this approach to calculate a novel relevance score that enables filtering of phenotypes and simplification of the topology.

5.2 Mapping Topic Model Results to HPO

Recent evaluation of HPO content coverage in UMLS showed that term coverage was only 54% [55]. Although additional HPO mappings were added in the 2015AB UMLS release [15], we still discovered that only 10-20% of topic results terms had mappings. To further improve the coverage of mappings from CUI to HPO, we utilized a deep learning approach to map text descriptions to HPO terms that is being developed in our lab [1], improving term coverage to 80-90%.

We manually audited the terms to ensure that the terms with higher probabilities ($>$0.001) were being correctly represented. To fix remaining errors, we introduced a curation layer to the mapping process that overrode results from the deep learning approach. The goal of this data mapping process was not to develop a generalized approached, but to ensure that our data was accurately represented using HPO terms. Efforts to improve mappings between medical vocabularies are a continuous effort and outside the scope of the present work.

The resulting data mapping process thus mapped individual CUI codes to HPO terms. These HPO terms were then used to extract relevant subgraphs of the HPO, using the method of our prior work [23, 24]. In so doing, probabilities for multiple CUI codes that mapped to the same HPO term were aggregated. Additionally, the subsumptive "is-a" relationships of the HPO supported logical inferences to higher-level phenotypes, enabling probabilities of specific phenotypes to be aggregated up the hierarchy to the root. The result is a multi-level description of increasing granularity, where more general phenotype terms combine the probabilities of more specific phenotype terms.

5.3 Computing Summary Measures

Based on the collaborative design process, we developed a set of measures to summarize the temporal evolution of each phenotype's probabilities to facilitate comparisons: maximum probability, trend slope, representative age interval, and relevance score. Without the graph-based data abstraction, it would not be possible to calculate temporal summary measures over the hierarchy of phenotypes.

We used two levels of age intervals to compute these summary measures: the overall time interval (0-14 years), as well as four discrete age intervals, which align to developmental milestones: *neonatal/infantile* (0-2 years), *childhood* (2-5 years), *early juvenile* (5-10 years), and *late juvenile* 10-14 years. These derive from the *Age of Onset* term in the HPO, although we subdivide juvenile into early and late in order to subdivide this longer interval.

5.3.1 Maximum Probability and Trend Slope

To summarize the estimated probabilities of each phenotype over patient age, we calculate two measures to summarize the overall time interval as a single value: the maximum probability and the slope of the linear regression fit. These statistics were chosen because they helped identify the most important probability and the magnitude of the change, respectively.

5.3.2 Representative Age Interval

To identify the age interval where the highest probability occurs, we compute the 95th percentile probability in each of the four discrete age intervals. The interval with the highest value is chosen as representative for a given phenotype. This coarse measure helps to localize the age interval where the peak probability of the phenotype occurs.

5.3.3 Relevance Score

To guide the interpretation of disease subtypes, we developed a novel measure of relevance for each phenotype to help identify phenotypes that are more likely to be unique to the modeled disease, regardless of the modeled probability. This relevance score improves upon the current approach of reducing the phenotype search space by considering only high-probability phenotypes, and was necessary due to the hierarchical aggregation of probabilities (i.e., general phenotypes will have higher probability than specific phenotypes).

The HPO has been used to calculate the frequency with which a phenotype is associated with known diseases [31]. We apply this computed frequency as an estimate of expected phenotype probability. Although this estimate is biased toward rare genetic diseases, representative prevalence data is difficult to source. Thus, this estimate was sufficient as a proof of concept to demonstrate the automatic calculation of phenotype relevance. The intuition follows that phenotypes with both a high expected probability and a high modeled probability are less unique, while phenotypes with a lower expected probability and a higher modeled probability are more likely to be unique to the disease.

To calculate the relevance score, we compute the difference between the probabilities of phenotypes in the topic model and the expected probabilities derived from the HPO. Let X be a phenotype in the topic model, then $P_{model}(X)$ is the phenotype probability in a topic at one time step and $P_{expected}(X)$ is the expected phenotype probability. We calculate the magnitude of the difference between $P_{model}(X)$ and $P_{expected}(X)$ using the signal to noise ratio:

$$SNR(X) = 10\log_{10}(P_{model}(X)) - 10\log_{10}(P_{expected}(X)) \quad (1)$$

The distribution of SNR is used to identify relevant phenotypes by assigning a percentile threshold, i.e., the phenotypes with larger positive differences in the right long-tail of the distribution indicate phenotypes with low expected and high modeled probability (Fig. 2).

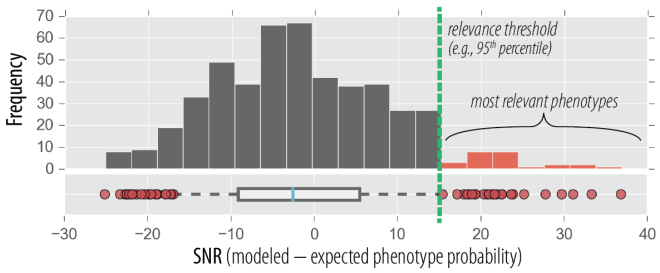

Fig. 2. An example SNR distribution for one topic at one time step. A percentile threshold can be used to identify the most relevant phenotypes in the right long-tail. The boxplot shows IQR and 5-95 percentiles.

6 PhenoLines Description

PhenoLines is an interactive visual analysis tool that was designed using the data transformation workflow. Three visual representations are used to summarize phenotypes within topics (i.e., radial hierarchy, scatterplot, rank-ordered list) and two to summarize phenotypes across topics (i.e., summary table, timeline charts) (Fig. 3).

To co-ordinate between these visualizations, PhenoLines supports fully-linked views. Phenotypes, represented as regions, points, or text, can be hovered-over or selected in the radial hierarchy, scatterplot, and rank-ordered lists to display detailed information using the summary table and timeline charts. Selecting a phenotype locks it as the default for detailed information in the Detail Panel (Fig. 1B). A keyword search supports the direct look-up of specific phenotype terms (Fig. 1D).

6.1 Compare Within Topics

The Topics Panel (Fig. 1C) provides an overview of all topics, organized into columns of visualizations for each topic. Three visualizations are used to describe each topic: a radial hierarchy, a scatterplot, and a rank-ordered list (Fig. 3C,D,E). These visualizations enable a user to explore the dominant phenotypes **within** each topic, supporting *browse* & *explore* search, through *identify* & *compare* queries.

We support these explorations through multiple visual representations (*encode*) that can be customized (*arrange* & *change*). The Settings Panel (Fig. 1A) enables the visual representations to be configured based on the analysis. For example, the user can change the summary measure used to define the fill color of all visual representations in the Topics Panel (Fig. 4). Color palettes were adapted from ColorBrewer [27]. The fill color thus explicitly encodes the temporal evolution of phenotype probabilities (i.e., maximum probability, trend slope, representative age interval).

The **radial hierarchy** represents the semantic relationships between phenotypes (Fig. 3C). General phenotypes are at the center (root), and grow in specificity toward the periphery (leafs). The user can set the arc length of each region to represent the maximum probability of each phenotype (Fig. 5A). This supports a top-down approach to phenotype exploration, whereby dominant high-level phenotypes can be identified, and then the contributing specific phenotypes can be investigated.

The **scatterplot** displays correlations between any of the four summary measures (Fig. 3D). This visualization enables a comparison of the distribution of phenotypes across the measures, with each phenotype represented as a data point in the scatterplot. The axes of the scatterplot can be customized to display different correlations (e.g., maximum probability v. representative age interval). In addition to the fill color setting, this enables the visual comparison of up to three summary measures at a time in the scatterplot.

The **rank-ordered list** summarizes the phenotypes with highest probability in each topic (Fig. 3E). The relative probability of each phenotype is communicated using bar charts adjacent to each term. The rank-ordered list supports a bottom-up approach to phenotype exploration, and the list can be filtered based on the height of each phenotype node in the graph. For example, the list can be limited to only leaf phenotypes (height=0), to include parents (height=1), or to include parents and grandparents (height=2). This enables the user to control the ratio of aggregated phenotypes to display along side the modeled phenotypes.

6.2 Compare Between Topics

The Topics Panel can also be used to make high-level comparisons between topics, i.e., *how are the phenotypes between topics different, overall?* The user can set the arc length to represent the descendant count of each phenotype. This supports comparisons of explicitly encoded values between topics through a consistent representation of the hierarchy topology, supporting *browse* & *explore* search, through *compare* & *summarize* queries (Fig. 5B).

A detailed comparison of a single phenotype across topics is facilitated through the Detail Panel (Fig. 1B), which contains a summary table and juxtaposed timeline charts for the phenotype in each topic. These visual representations support *locate* & *explore* search, through *identify* & *compare* queries.

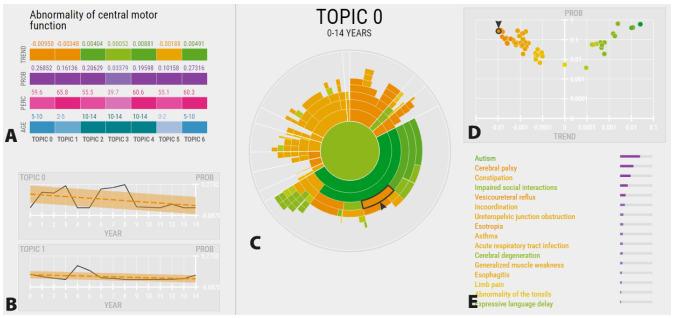

Fig. 3. Phenotypes can be compared across topics (A) summary table and (B) juxtaposed timeline charts, and also within topics (C) radial hierarchy, (D) scatterplot, and (E) rank-ordered list.

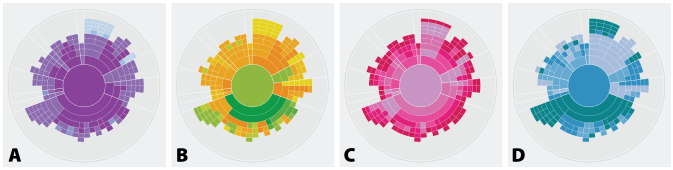

Fig. 4. The visualizations can encode different summary measures. Example topic displaying (A) maximum probability, (B) trend slope, (C) relevance score percentile, and (D) representative age interval.

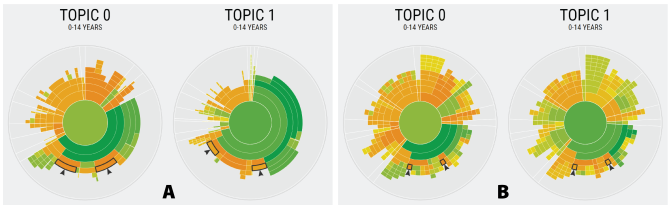

Fig. 5. Arc length based on (A) maximum probability reveals dominant phenotypes, or (B) descendant count facilitates value comparisons between topics through consistent topology. The same two topics are depicted in both panels; only arc length is changed.

The **summary table** displays the name of the phenotype alongside a table of all summary measures, enabling for the detailed comparison of a single phenotype across all topics (Fig. 3A).

A **timeline chart** for each topic displays how the probability of the phenotype changes with a patient's age. These timeline charts are vertically juxtaposed to facilitate comparisons along the temporal dimension of the evolving phenotype probabilities (Fig. 3B).

The Detail Panel complements the Topics Panel: once a phenotype is identified in the topic charts, the phenotype details can be carefully inspected and compared across topics and time.

6.3 Filtering Phenotypes and Compressing Hierarchies

To address visual complexity in all visual representations, we use *filter* & *aggregate* approaches to simplify the visualizations. As the topology of the HPO is complex and supports multiple inheritance, it can be difficult to fully display [23, 24]. During the Design Stage, it became apparent that it was unnecessary to show the entire hierarchy all the time. However, a data-aware approach to non-uniform hierarchy pruning was required to ensure important phenotypes were not removed. The graph-based data abstraction and relevance score made filtering and topology simplification possible, since it allowed systematic traversals of the phenotype relations and recursive application of relevance thresholds. As filtering and simplification is carried out on the graph representation, the effects propagate to all visual representations, simplifying the radial hierarchy, reducing occlusions in the scatterplot, and removing phenotypes from the rank-ordered list.

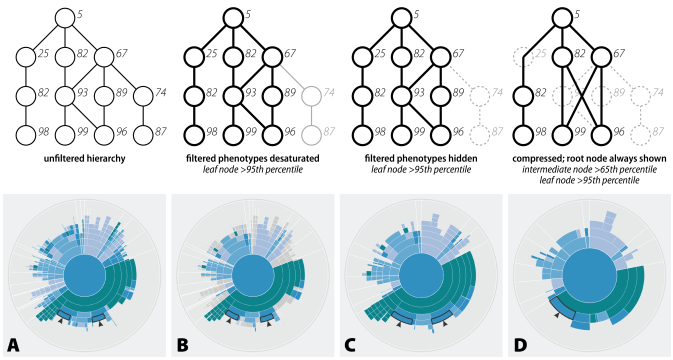

Fig. 6. Phenotypes can be filtered or compressed using a percentile threshold on the relevance score distribution. (A) Unfiltered hierarchy. (B) Filtered phenotypes can be desaturated, or (C) hidden to reduce breadth. (D) Compression uses two thresholds to eliminate intermediate phenotypes to reduce depth. An example radial hierarchy is shown here, but filtering and compression affects all visual representations.

The relevance score is exposed to users as a percentile of the SNR distribution, and is used to filter less relevant phenotypes by defining a minimum relevance percentile threshold. Our approach adopted the Minimum Description Length (MDL) Treecuts method to prune branches of the hierarchy [52]. MDL Treecuts evaluate the nodes of the hierarchy recursively to define non-uniform treecuts based on a measure of importance—in our implementation we use the relevance score. Thus, branches of the hierarchy are filtered if they do not contain at least one descendant with a relevance score percentile higher than the threshold. The user can dynamically set the minimum relevance percentile threshold to define the aggressiveness of filtering. For example, phenotypes can be filtered to only include those in the top 10% of relevance scores by setting the threshold to the 90th percentile. The phenotypes below this threshold are either displayed using a desaturated scale (Fig. 6B) or can be hidden from view (Fig. 6C).

Although filtering removes branches from the hierarchy, it does not reduce depth when leaf phenotypes are also the most relevant. To address this, we developed a hierarchy compression method, also based on the relevance score. A secondary percentile threshold is used to define the minimum relevance score percentile required to display an intermediate phenotype. Intermediate phenotypes below this threshold are hidden and descendant phenotypes are merged to reduce the depth of the hierarchy. A combination of the filtering and compression methods results in a simplified hierarchical topology that provides higher-level context, while retaining relevant leaf phenotypes (Fig. 6D).

7 Initial Evaluations

We conducted initial evaluations of PhenoLines in two parts. First, to investigate debugging and refining topic models. Second, to observe usage during a characterization session with a developmental-behavioral pediatrician. Three topic models were generated for the evaluations.

Baseline Model. The Baseline Model was trained using a non-informative Dirichlet prior based on an equal likelihood that a phenotype would appear in any topic. This approach assumes no prior knowledge about the topics and is often used for initial topic discovery.

Informed Prior Model. The Informed Prior Model used a Dirichlet prior derived from an empirical optimization that maximized the likelihood of phenotypes based on their occurrence in the original data. This symmetric prior promoted the probability of phenotypes which were known to occur with higher frequency.

Curated Model. The Curated Model used a constrained dataset, limited to phenotypes that occurred with high probability in the Baseline Model. Using unsupervised learning, an initial set of terms with high probability was expanded to include related terms. The resulting set of terms included 246 phenotypes. The Curated Model was then trained on this data. This approach removed less relevant terms from the dataset to reduce the noise and improve differentiation between topics.

7.1 Debugging and Refining

Visualizing the hierarchical and temporal components of the topics using PhenoLines could help machine learning researchers interpret topic models for the purposes of debugging and refinement. To investigate this process, we engaged our machine learning collaborators to use PhenoLines to identify and explain the differences between the outputs of the Baseline Model, the Informed Prior Model, and the Curated Model. We elicited additional feedback from two third-party machine learning experts through a short-term deployment study to corroborate our findings.

7.1.1 Informed Prior vs. Baseline

The topics produced by the Informed Prior Model were contrasted against those of the Baseline Model. Although the expectation was that the informed prior would yield superior results, a precise understanding of how the topics improved is needed to guide future refinements.

Our collaborators used PhenoLines to compare and contrast the topics produced by the Informed Prior Model and the Baseline Model. In both models, there was little differentiation between topics for specific phenotypes with high probability, but substantial differentiation when many related, but low probability phenotypes were aggregated. By enabling investigations of general phenotypes (i.e., biological systems), the visualizations helped our collaborators to explain how the symmetric prior improved topic differentiation. Using a combination of the radial hierarchies and the timeline charts, specific differences between the topics could be articulated. For example, two topics having high prevalence of gastrointestinal abnormalities were distinguished by their temporal characteristics: one was strongly increasing over age, while the other was decreasing.

7.1.2 Curated vs. Informed Prior

The insights obtained from the Informed Prior Model suggested the approach of constraining terms, to filter terms common across the population and thereby further promote topic differentiation.

The topics produced by the Curated Model were better differentiated than the Informed Prior Model, and these differentiations extended to specific phenotypes in three of the topics. Our collaborators again used the radial hierarchies to identify these differences. For example, two topics were enriched with neurological abnormalities, but these were differentiated by more specific phenotypes that indicated prevalence of mood disorders in one, and psychosis in the other. The radial hierarchy and timeline charts helped to characterize the differences between the topics and identify patterns that could be validated with medical experts.

The insights of how topics differentiated in each of the three models would not have been possible without the hierarchical aggregation of phenotype probabilities. Had the topics been compared only using the highest probability phenotypes, the differentiation between the topics would have been missed.

7.1.3 Third-Party Evaluation

To corroborate our findings and collect qualitative feedback, we recruited two experienced third-party machine learning experts to evaluate PhenoLines over a one-week deployment. Both had experience with topic modeling; one had prior experience with ASD. The tool was demonstrated in a 1 hour session, in which experts were trained in the interface and the three topic models were introduced. Experts were asked to use the tool and complete an online survey to document their hypotheses and findings after each usage. Each conducted 2-3 analysis sessions, lasting 45 minutes to 1 hour each. A 30 minute semi-structured interview was conducted after the deployment to collect detailed feedback, using the survey results to guide the discussion.

Both experts found that the topics produced by the Curated Model were more discriminative than those produced by the other two models, and spent most of their time investigating this model. The expert with no ASD experience identified the same three differentiated topics as our collaborators. The expert with ASD experience also identified these topics and commented that he was able to distinguish comorbidities with which he was familiar. Knowledge of ASD impacted how these experts used PhenoLines. The rank-ordered list was used in all sessions

by the expert familiar with ASD, while the expert without experience started using it as a reference in later sessions. He explained that it became very useful after he had familiarized himself with the terms, and that he used the radial hierarchy to learn about the terms and their relationships. Both experts reported that the radial hierarchies and timeline charts were the most informative when differentiating topics, while the scatterplot helped to locate terms with similar summary measures regardless of their hierarchical and temporal characteristics.

The experts also suggested improvements. Both wanted additional information from the training data included, such as which CUI codes were mapped to each phenotype, and the number of patients that were modeled by each topic. The expert with ASD experience also wanted to know to what CUI codes the aggregated phenotypes were related.

7.2 Characterization and Validation

PhenoLines was evaluated in situ, during a topic characterization session between a machine learning researcher and a developmental-behavioral pediatrician. The session lasted two hours and topics from the *Curated Model* were used. The pediatrician had prior experience reviewing disease clustering results with the machine learning researcher, but no prior experience with PhenoLines or topic models. In the first 30 minutes, the machine learning researcher introduced PhenoLines to the pediatrician and demonstrated how aspects of the topic model were visualized. Over the next 60 minutes, the two collaboratively used PhenoLines to characterize different topics in the model. The final 30 minutes were used for open discussion to collect qualitative feedback about PhenoLines.

7.2.1 Topic Characterization Results

Of the seven topics in the model, three were characterized as likely to align with cases the pediatrician was familiar with: (a) children with a diagnosis of cerebral palsy and ASD, (b) children with an earlier misdiagnosis of ASD and a later correct diagnosis of ADHD and mood disorders, and (c) children with an earlier misdiagnosis of ASD and a later correct diagnosis of psychosis. Another topic was characterized as a likely catch-all for phenotypes that did not directly relate to the diagnosis of ASD. The remaining four topics appeared to characterize diagnoses of ASD occurring at different ages and were differentiated along comorbidities that the pediatrician was unsure related to ASD.

These results are in line with prior clustering results using this dataset [16], in which subgroups were identified with primarily neurological (e.g., seizures), psychiatric (e.g., mood disorders), multi-system, and undefined characterizations. In addition to replicating these results, PhenoLines enabled a richer discussion of the specific phenotypes in each topic. For example, seizures and contractures are highly correlated with cerebral palsy and this relationship was evident in the topic having to do with cerebral palsy and ASD. These additional details provided stronger evidence to support the topic characterizations. The hierarchical aggregation of probabilities also made the contributions of specific phenotypes to higher-level phenotypes more apparent.

The pediatrician was particularly interested in the two topics that appeared to characterize misdiagnoses of ASD. She commented that specific patient cases of misdiagnosis are very hard to identify in practice, but are of significant interest to the clinical community. Early correct diagnosis of mood disorders and psychosis is difficult and a misdiagnosis of ASD often occurs in such cases. She commented that there was great value in using the topics to characterize comorbid phenotypes that could potentially help differentiate these patients earlier and lead to more robust diagnosis strategies.

7.2.2 Differentiated Roles

An interesting dynamic evolved between the pediatrician and the machine learning researcher, whereby the pediatrician took on the role of investigator and the machine learning researcher the role of verifier. The pediatrician focused primarily on the rank-ordered list. She explained it was the easiest for her to interpret because she quickly recognized relevant phenotype terms and efficiently parsed the relative probabilities that were represented as bars. Since doctors are trained to think in terms of specific symptoms, it is not surprising that the rank-ordered list most aligned to her training. However, the rank-ordered list also hides the complexities of hierarchical relationships and temporal trends, and could lead to misinterpretations based on an incomplete reading of the data. Thus, as the pediatrician developed hypotheses, the machine learning researcher ensured that the hypotheses were grounded in the data. Using the radial hierarchies, timeline charts, and scatterplots, the machine learning researcher collected additional evidence to support or counter the hypotheses developed by the pediatrician in this verification process. The machine learning researcher commented that without the holistic perspective provided by multiple visual representations, this verification task would not have been possible.

The two participants also reflected on differences compared to prior characterization sessions. In the past, sessions were conducted using prepared lists and static charts. Although similar representations were available in PhenoLines, both experts agreed that the interactivity of the tool greatly improved the flow of the discussion. The interactivity enabled immediate inspection of any part of the model, whereas in the past, the machine learning researcher would have had to follow-up regarding phenotypes for which no materials had been prepared. Using PhenoLines, the discussion was more fluid and allowed a deeper line of investigation to occur without interruption.

The machine learning researcher commented that PhenoLines was integral in facilitating the dialogue with the pediatrician. By supporting a differentiation of roles, the experts were able to simultaneously attend to different visual representations and collaboratively develop robust explanations of the relationships captured by the topics.

7.2.3 Preferred Visual Representations

The rank-ordered list, with the representative age interval color scale, was used most by the pediatrician and preferred. It enabled her to quickly draw coarse associations between high probability phenotypes without necessitating interpretation of the more complex visual representations. This combination enabled the pediatrician to speculate about the topics that characterized misdiagnoses.

On the other hand, the machine learning researcher reported that the radial hierarchies was most preferred. When preparing for the characterization session, the hierarchy enabled a top-down approach, first identifying the dominant biological systems represented in a topic and then drilling down to the more specific contributing pheotypes. In contrast, the radial hierarchies were used in a bottom-up manner during the characterization session. As the pediatrician focused on detailed phenotypes, the machine learning researcher used the hierarchies to verify whether the trends of the specific phenotypes that the pediatrician identified were mirrored in related specific phenotypes or in more general phenotypes. This verification process proved crucial to contextualize the topic characterizations.

The timeline charts were reported most useful to compare the specific age intervals where a phenotype dominated, by looking at the distribution of probabilities between topics, and to verify the ages with the most probability mass. This helped to differentiate topics that captured observations of disease symptoms at an earlier or later age. Although the representative age interval measure provided a coarse indication of the age, the timeline charts were critical in verifying the actual distribution and magnitude of the probability. The machine learning researcher commented this workflow was efficient when identifying phenotypes for further investigation. When applying the representative age interval to the radial hierarchies, judgements could be made regarding whether related phenotypes presented at similar age intervals, or whether there was a discrepancy.

Two novel ad hoc usages of the scatterplots were observed. First, the scatterplots were preferred by both the machine learning researcher and the pediatrician to compare the maximum probability of phenotypes. By encoding maximum probability along the vertical axis they could accurately and simultaneously compare phenotype probabilities, not only within topics, but also across the topics. Second, the scatterplots were used to perform A/B comparisons, by selecting a phenotype and then quickly hovering over other phenotypes to flip the information displayed in the Details Panel. These observations can inform improvements to the interface to address these analysis use cases.

7.2.4 Relevance Score

The pediatrician found the relevance score was effective at filtering less relevant phenotypes, but noted there were several errors. Further investigation revealed these occurred due to biases present in the expected phenotype probability. As the expected probability derived from the HPO is biased toward rare genetic diseases, the annotations have a higher prevalence of neurological and developmental abnormalities, (e.g., developmental delay, intellectual disability, seizures). As these abnormalities are overrepresented in the expected probability, the associated modeled phenotypes received lower relevance scores. These abnormalities are highly relevant to ASD and uncommon in most non-genetic conditions. That said, she noted that automatically quantifying the relevance of a phenotype is highly desirable, since it can help identify meaningful phenotypes regardless of their modeled probability. Thus our approach is promising, but requires a curated expected probability to avoid misleading interpretations.

8 Discussion and Limitations

We addressed the top three levels of the Nested Model for Design and Validation [40]. The visual encodings were justified using collaborative design, and validated, along with the data abstraction, via evaluations with machine learning experts. The problem domain characterization was validated through observation of a disease subtype characterization session involving a machine learning researcher and a medical expert.

Phenotype hierarchies. The machine learning researchers unanimously lauded the representation of phenotypes within the hierarchical taxonomy of the HPO. The aggregation of phenotype probabilities helped identify the dominant biological systems in each topic and acted as a starting point for deeper investigation. Although the pediatrician reported some of the HPO terminology was unfamiliar (when compared with ICD terms), the radial hierarchy enabled discussions about biological systems that are not available as CUI codes. The hierarchical representation also promoted investigations of specific phenotype divergences between topics with similar dominant biological systems, which was not possible with previous analysis approaches.

Temporal evolution of phenotype probabilities. The timeline charts were used extensively by both the machine learning experts and the medical expert to investigate the evolution of specific phenotype probabilities. Timeline charts were used most frequently to confirm patterns observed in the radial hierarchies, and explore the implications by comparing the timeline charts of a specific phenotype across the topics. This division of hierarchical and temporal perspectives aligned well with the workflow of experts in both domains.

Temporal summary measures. The machine learning experts reported the maximum probability and relevance score percentiles aligned with the goals of identifying dominant phenotypes in each topic and intuiting about their relevance to the disease. The trend slope and representative age interval provided meaningful information when investigations focused on biological systems where probabilities were aggregated. However, when investigating specific phenotypes, the modeled probabilities were often isolated to shorter time intervals, so the trend slope over the entire age interval did not summarize these distributions well. In response, we introduced an additional option to calculate the trend slope for only the representative age interval. There would be value in further efforts to develop a metric that accurately identifies the time interval of the largest probability mass.

Facilitating comparisons. The visualizations supported a variety of comparisons of the topics: hierarchical relationships within topics, temporal progression across topics, and correlations between summary measures. The ability to highlight and select phenotypes in any view was greatly appreciated, as it enabled for the isolation of specific terms to compare and provided a means of ad hoc A/B comparisons between phenotypes. An additional feature that was requested was the ability to hide phenotypes that are known to be less relevant.

Expected Phenotype Probability. We identified shortcomings of the relevance score due to the expected probability based on the HPO. Developing expected probability is challenging, because finding a representative dataset to model is difficult [30]. Patients who frequently visit hospitals have more data entered, leading to an over-representation of individuals with pre-existing conditions. Hospitals also have different specializations and tend to see patients with associated conditions. Data from an institution may thus also have a bias toward patients with certain types of conditions (e.g., gastrointestinal), which could lead to an over-representation of these phenotypes. In any case, the approach of calculating a relevance score based on an expected probability is promising, with the caveat that expected probabilities must be carefully curated and validated to ensure that they do not mislead the investigators. This is interesting and highly important future work to make it easier for medical experts to reason about the complex output of topic models.

Coupling to clinical data. The pediatrician suggested PhenoLines could provide a starting point for clinical investigations, based on topics that characterize interesting disease processes or patient cohorts. Providing a tighter coupling to the underlying EHR data could extend the clinical utility of PhenoLines. The phenotype terminology used in the HPO is less familiar to medical experts than ICD code descriptions, so integrating term mappings between other medical terminologies could improve the interpretability of the visualizations.

9 Conclusion and Future Work

This work introduced PhenoLines, an interactive visual analysis tool to support the interpretation of disease subtypes via topic models to facilitate model characterization and optimization. We described a data transformation workflow to produce a flexible graph-based data abstraction that can be converted into a variety of visual representations that provide complementary perspectives on topics and the phenotypes they model. We demonstrated the benefits of both hierarchical and temporal representations for the analysis of disease subtypes. Results of initial evaluations suggest that PhenoLines aids interpretation of the quality of topic models by machine learning researchers, and also enables fluid collaborative inspection of topic models with medical experts. Improving our understanding of how phenotypes manifest and evolve over time for a particular disease can help differentiate disease subtypes, improving diagnosis and prognosis, and enabling for more effective personalized care to address the individual needs of patients.

Feedback suggested that PhenoLines could act as a gateway to the underlying EHR data. For example, topics identified as characterizing cases of misdiagnosis could be used to extract patient cohorts for further clinical investigation. Extending the clinical utility of PhenoLines is an interesting direction for future research.

While the relevance score is a promising approach to simplify the hierarchical topology through filtering and compression, further research is necessary to develop and validate representative expected phenotype probabilities to improve automatic relevance computation.

In our evaluations, the topic models were limited to seven topics, but future work could address interactive adjustment of the number of topics. Methods that automatically derive the optimal number of topics may be applicable (e.g., Hierarchical Dirichlet Process [51]).

Although our work investigated disease characterization via topic models, the data transformation workflow is general and could be applied to other ontologies, and for data with or without a temporal component. Just as the ontology-based data abstraction provided common ground for discussions between machine learning researchers and medical experts, such abstractions could help bridge the knowledge-gap between collaborating visualization researchers and domain experts. A graph-based abstraction is familiar in the visualization domain, while the terminology aligns to the mental models of domain experts. As we demonstrated, the graph-based data abstraction can be flexibly transformed into a variety of visual representations. Further, our approach to topology simplification could be adapted to general graphs or trees when expected measures of data are available.

Acknowledgments

This work was partially funded by Genome Canada and Ontario Genomics through a Bioinformatics/Computational Biology grant to Dr. Brudno. The authors thank Michelle Annett, Aryan Arbabi, Rafael Veras, Bruno De Araujo, John Hancock, the domain experts, as well as the anonymous reviewers for their thoughtful suggestions.

REFERENCES

[1] A. Arbabi. Personal communication, 2016-11-03.

[2] D. M. Blei and J. D. Lafferty. Dynamic topic models. In *Proceedings of the International Conference on Machine Learning*, pages 113–120. ACM, 2006.

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3(Jan):993–1022, 2003.

[4] O. Bodenreider. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, 32(suppl 1):D267–D270, 2004.

[5] M. Bostock, V. Ogievetsky, and J. Heer. D^3 data-driven documents. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2301–2309, 2011.

[6] M. Brehmer and T. Munzner. A multi-level typology of abstract visualization tasks. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2376–2385, 2013.

[7] J. Chang, J. L. Boyd-Graber, S. Gerrish, C. Wang, and D. M. Blei. Reading tea leaves: How humans interpret topic models. In *Proceedings of the Conference on Neural Information Processing Systems*, volume 31, pages 1–9, 2009.

[8] Z. Che, D. Kale, W. Li, M. T. Bahadori, and Y. Liu. Deep computational phenotyping. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 507–516. ACM, 2015.

[9] Y. Chen, L. Wang, M. Dong, and J. Hua. Exemplar-based visualization of large document corpus. *IEEE Transactions on Visualization and Computer Graphics*, 15(6), 2009.

[10] L. Chittaro, C. Combi, and G. Trapasso. Data mining on temporal data: a visual approach and its clinical application to hemodialysis. *Journal of Visual Languages & Computing*, 14(6):591–620, 2003.

[11] J. Chuang, S. Gupta, C. D. Manning, and J. Heer. Topic model diagnostics: Assessing domain relevance via topical alignment. In *Proceedings of the International Conference on Machine Learning*, pages 612–620, 2013.

[12] J. Chuang, C. D. Manning, and J. Heer. Termite: Visualization techniques for assessing textual topic models. In *Proceedings of the International Working Conference on Advanced Visual Interfaces*, pages 74–77. ACM, 2012.

[13] W. Cui, S. Liu, L. Tan, C. Shi, Y. Song, Z. Gao, H. Qu, and X. Tong. TextFlow: Towards better understanding of evolving topics in text. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2412–2421, 2011.

[14] W. Cui, S. Liu, Z. Wu, and H. Wei. How hierarchical topics evolve in large text corpora. *IEEE Transactions on Visualization and Computer Graphics*, Volume: 20, Issue: 12:2281 – 2290, November 2014.

[15] F. Dhombres and O. Bodenreider. Interoperability between phenotypes in research and healthcare terminologies – investigating partial mappings between HPO and SNOMED CT. *Journal of Biomedical Semantics*, 7(1):3, 2016.

[16] F. Doshi-Velez, Y. Ge, and I. Kohane. Comorbidity clusters in autism spectrum disorders: an electronic health record time-series analysis. *Pediatrics*, 133(1):e54–e63, 2014.

[17] F. Doshi-Velez, B. C. Wallace, and R. Adams. Graph-sparse lda: a topic model with structured sparsity. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

[18] W. Dou, X. Wang, R. Chang, and W. Ribarsky. ParallelTopics: A probabilistic approach to exploring document collections. In *Proceedings of the IEEE Conference on Visual Analytics Science and Technology*, pages 231–240. IEEE, 2011.

[19] W. Dou, L. Yu, X. Wang, Z. Ma, and W. Ribarsky. HierarchicalTopics: Visually exploring large text collections using topic hierarchies. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2002–2011, 2013.

[20] H. M. Elibol, V. Nguyen, S. Linderman, M. Johnson, A. Hashmi, and F. Doshi-Velez. Cross-corpora unsupervised learning of trajectories in autism spectrum disorders. *Journal of Machine Learning Research*, 17(133):1–38, 2016.

[21] A. Ganesan, K. Brantley, S. Pan, and J. Chen. LDAExplore: Visualizing topic models generated using latent dirichlet allocation. *CoRR*, abs/1507.06593, 2015.

[22] M. J. Gardner, J. Lutes, J. Lund, J. Hansen, D. Walker, E. Ringger, and K. Seppi. The Topic Browser: An interactive tool for browsing topic models. In *Conference on Neural Information Processing Systems, Workshop on Challenges of Data Visualization*, volume 2, 2010.

[23] M. Glueck, A. Gvozdik, F. Chevalier, A. Khan, M. Brudno, and D. Wigdor. PhenoStacks: cross-sectional cohort phenotype comparison visualizations. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):191–200, 2017.

[24] M. Glueck, P. Hamilton, F. Chevalier, S. Breslav, A. Khan, D. Wigdor, and M. Brudno. PhenoBlocks: Phenotype comparison visualizations. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):101–110, 2016.

[25] B. Gretarsson, J. Odonovan, S. Bostandjiev, T. Höllerer, A. Asuncion, D. Newman, and P. Smyth. TopicNets: Visual analysis of large text corpora with topic modeling. *ACM Transactions on Intelligent Systems and Technology*, 3(2):23, 2012.

[26] T. Gschwandtner, W. Aigner, K. Kaiser, S. Miksch, and A. Seyfang. CareCruiser: exploring and visualizing plans, events, and effects interactively. In *2011 IEEE Pacific Visualization Symposium*, pages 43–50. IEEE, 2011.

[27] M. Harrower and C. A. Brewer. ColorBrewer.org: an online tool for selecting colour schemes for maps. *The Cartographic Journal*, 40(1):27–37, 2003.

[28] V. Huddar, B. K. Desiraju, V. Rajan, S. Bhattacharya, S. Roy, and C. K. Reddy. Predicting complications in critical care using heterogeneous clinical data. *IEEE Access*, 4:7988–8001, 2016.

[29] T. Iwata, T. Yamada, and N. Ueda. Probabilistic latent semantic visualization: topic model for visualizing documents. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 363–371. ACM, 2008.

[30] I. S. Kohane, D. R. Masys, and R. B. Altman. The incidentalome: a threat to genomic medicine. *Journal of the American Medical Association*, 296(2):212–215, 2006.

[31] S. Köhler, M. H. Schulz, P. Krawitz, S. Bauer, S. Dölken, C. E. Ott, C. Mundlos, D. Horn, S. Mundlos, and P. N. Robinson. Clinical diagnostics in human genetics with semantic similarity searches in ontologies. *The American Journal of Human Genetics*, 85(4):457–464, 2009.

[32] S. Köhler, N. A. Vasilevsky, M. Engelstad, E. Foster, J. McMurry, S. Aymé, G. Baynam, S. M. Bello, C. F. Boerkoel, K. M. Boycott, et al. The human phenotype ontology in 2017. *Nucleic Acids Research*, page gkw1039, 2016.

[33] J. Krause, A. Perer, and H. Stavropoulos. Supporting iterative cohort construction with visual temporal queries. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):91–100, 2016.

[34] H. Lee, J. Kihm, J. Choo, J. Stasko, and H. Park. iVisClustering: An interactive visual document clustering via topic modeling. In *Computer Graphics Forum*, volume 31, pages 1155–1164. Wiley Online Library, 2012.

[35] T. Lingren, P. Chen, J. Bochenek, F. Doshi-Velez, P. Manning-Courtney, J. Bickel, L. W. Welchons, J. Reinhold, N. Bing, Y. Ni, et al. Electronic health record based algorithm to identify patients with autism spectrum disorder. *PloS one*, 11(7):e0159621, 2016.

[36] S. Liu, X. Wang, J. Chen, J. Zhu, and B. Guo. TopicPanorama: A full picture of relevant topics. In *Proceedings of the IEEE Conference on Visual Analytics Science and Technology*, pages 183–192. IEEE, 2014.

[37] S. Liu, M. X. Zhou, S. Pan, Y. Song, W. Qian, W. Cai, and X. Lian. Tiara: Interactive, topic-based visual text summarization and analysis. *ACM Transactions on Intelligent Systems and Technology*, 3(2):25, 2012.

[38] S. Malik, F. Du, M. Monroe, E. Onukwugha, C. Plaisant, and B. Shneiderman. Cohort comparison of event sequences with balanced integration of visual analytics and statistics. In *Proceedings of the 20th International Conference on Intelligent User Interfaces*, pages 38–49. ACM, 2015.

[39] J. Munkres. Algorithms for the assignment and transportation problems. *Journal of the Society for Industrial and Applied Mathematics*, 5(1):32–38, 1957.

[40] T. Munzner. A nested model for visualization design and validation. *IEEE Transactions on Visualization and Computer Graphics*, 15(6), 2009.

[41] D. Newman, T. Baldwin, L. Cavedon, E. Huang, S. Karimi, D. Martinez, F. Scholer, and J. Zobel. Visualizing search results and document collections using topic maps. *Web Semantics: Science, Services and Agents on the World Wide Web*, 8(2):169–175, 2010.

[42] R. Pivovarov, A. J. Perotte, E. Grave, J. Angiolillo, C. H. Wiggins, and N. Elhadad. Learning probabilistic phenotypes from heterogeneous EHR data. *Journal of Biomedical Informatics*, 58:156–165, 2015.

[43] C. Plaisant, R. Mushlin, A. Snyder, J. Li, D. Heller, and B. Shneiderman.

LifeLines: using visualization to enhance navigation and analysis of patient records. In *Proceedings of the American Medical Informatics Association Symposium*, page 76. American Medical Informatics Association, 1998.

[44] R. Řehůřek and P. Sojka. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, 2010.

[45] A. Rind, W. Aigner, S. Miksch, S. Wiltner, M. Pohl, T. Turic, and F. Drexler. Visual exploration of time-oriented patient data for chronic diseases: Design study and evaluation. In *Symposium of the Austrian HCI and Usability Engineering Group*, pages 301–320. Springer, 2011.

[46] A. Rind, T. D. Wang, W. Aigner, S. Miksch, K. Wongsuphasawat, C. Plaisant, and B. Shneiderman. Interactive information visualization to explore and query electronic health records. *Foundations and Trends in Human-Computer Interaction*, 5(3):207–298, 2011.

[47] P. N. Robinson, C. J. Mungall, and M. Haendel. Capturing phenotypes for precision medicine. *Molecular Case Studies*, 1(1):a000372, 2015.

[48] M. Sedlmair, M. Meyer, and T. Munzner. Design study methodology: Reflections from the trenches and the stacks. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2431–2440, 2012.

[49] B. Shneiderman, C. Plaisant, and B. W. Hesse. Improving healthcare with interactive visualization. *Computer*, 46(5):58–66, 2013.

[50] C. Sievert and K. E. Shirley. LDAvis: A method for visualizing and interpreting topics. In *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*, pages 63–70, 2014.

[51] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Sharing clusters among related groups: Hierarchical dirichlet processes. In *Advances in Neural Information Processing Systems*, pages 1385–1392, 2005.

[52] R. Veras and C. Collins. Optimizing hierarchical visualizations with the minimum description length principle. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):631–640, 2017.

[53] T. D. Wang, C. Plaisant, A. J. Quinn, R. Stanchak, S. Murphy, and B. Shneiderman. Aligning temporal data by sentinel events: discovering patterns in electronic health records. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 457–466. ACM, 2008.

[54] X. Wang and A. McCallum. Topics over time: a non-markov continuous-time model of topical trends. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 424–433. ACM, 2006.

[55] R. Winnenburg and O. Bodenreider. Coverage of phenotypes in standard terminologies. *Joint Bio-Ontologies and BioLINK ISMB*, pages 41–44, 2014.

[56] K. Wongsuphasawat, J. A. Guerra Gómez, C. Plaisant, T. D. Wang, M. Taieb-Maimon, and B. Shneiderman. LifeFlow: visualizing an overview of event sequences. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1747–1756. ACM, 2011.

[57] Z. Zhang, D. Gotz, and A. Perer. Iterative cohort analysis and exploration. *Information Visualization*, 14(4):289–307, 2015.