JAMIA Open, 4(3), 2021, 1–8
doi: 10.1093/jamiaopen/ooab084
Research and Applications

Research and Applications

A natural language processing pipeline to synthesize patient-generated notes toward improving remote care and chronic disease management: a cystic fibrosis case study

Syed-Amad Hussain[1], Emre Sezgin [1], Katelyn Krivchenia[2,3], John Luna[1], Steve Rust[1], and Yungui Huang[1]

[1]IT Research and Innovation, The Abigail Wexner Research Institute, Nationwide Children's Hospital, Columbus, Ohio, USA, [2]Department of Pulmonary Medicine, Nationwide Children's Hospital, Columbus, Ohio, USA, and [3]Department of Pediatrics, The Ohio State University College of Medicine, Columbus, Ohio, USA

Corresponding Author: Emre Sezgin, PhD, The Abigail Wexner Research Institute, Nationwide Children's Hospital, 700 Children's Drive, Columbus, OH 43205, USA; emre.sezgin@nationwidechildrens.org

Received 14 July 2021; Revised 8 September 2021; Editorial Decision 12 September 2021; Accepted 14 September 2021

ABSTRACT

Objectives: Patient-generated health data (PGHD) are important for tracking and monitoring out of clinic health events and supporting shared clinical decisions. Unstructured text as PGHD (eg, medical diary notes and transcriptions) may encapsulate rich information through narratives which can be critical to better understand a patient's condition. We propose a natural language processing (NLP) supported data synthesis pipeline for unstructured PGHD, focusing on children with special healthcare needs (CSHCN), and demonstrate it with a case study on cystic fibrosis (CF).

Materials and Methods: The proposed unstructured data synthesis and information extraction pipeline extract a broad range of health information by combining rule-based approaches with pretrained deep-learning models. Particularly, we build upon the scispaCy biomedical model suite, leveraging its named entity recognition capabilities to identify and link clinically relevant entities to established ontologies such as Systematized Nomenclature of Medicine (SNOMED) and RXNORM. We then use scispaCy's syntax (grammar) parsing tools to retrieve phrases associated with the entities in medication, dose, therapies, symptoms, bowel movements, and nutrition ontological categories. The pipeline is illustrated and tested with simulated CF patient notes.

Results: The proposed hybrid deep-learning rule-based approach can operate over a variety of natural language note types and allow customization for a given patient or cohort. Viable information was successfully extracted from simulated CF notes. This hybrid pipeline is robust to misspellings and varied word representations and can be tailored to accommodate the needs of a specific patient, cohort, or clinician.

Discussion: The NLP pipeline can extract predefined or ontology-based entities from free-text PGHD, aiming to facilitate remote care and improve chronic disease management. Our implementation makes use of open source models, allowing for this solution to be easily replicated and integrated in different health systems. Outside of the clinic, the use of the NLP pipeline may increase the amount of clinical data recorded by families of CSHCN and ease the process to identify health events from the notes. Similarly, care coordinators, nurses and clinicians would be able to track adherence with medications, identify symptoms, and effectively intervene to improve clinical care. Furthermore, visualization tools can be applied to digest the structured data produced by the pipeline in support of the decision-making process for a patient, caregiver, or provider.

Conclusion: Our study demonstrated that an NLP pipeline can be used to create an automated analysis and reporting mechanism for unstructured PGHD. Further studies are suggested with real-world data to assess pipeline performance and further implications.

Key words: natural language processing, chronic disease, cystic fibrosis, patient notes, artificial intelligence

LAY SUMMARY

Free-text (or unstructured) patient notes are fundamental components for understanding patient's health conditions outside of the hospital, and therefore, impact healthcare and clinical decisions. These patient notes could be created through medical diaries, mobile apps, and devices through typing or speech transcriptions. We proposed a natural language processing model to analyze patient notes and extract critical information to improve the understanding of patient notes. To present the model, we used a cystic fibrosis case study and simulated patient notes. Our model was able to retrieve phrases associated with medication, dose, therapies, symptoms, bowel movements, and nutrition information. The model could be embedded to mobile apps or web portals to analyze patient notes and timely inform patients and caregivers. Furthermore, integration with to electronic medical records could enable providers to timely access patient health information and improve shared decision-making.

INTRODUCTION

In today's healthcare system, a large volume of patient health information is stored by healthcare providers (HCP) within their electronic health records (EHR) systems. However, much of the information is dependent on a patient's recall of personal health events occurring outside the clinic (symptoms, medication compliance, over-the-counter medicines, etc.), and the HCP's interpretation before recording these events in the EHR. This approach may lead to noncohesive, incomplete, and even erroneous health records,[1,2] which may limit providers' understanding of a patient's condition and negatively impact clinical decision-making.[3] This issue raises concerns for patients with chronic conditions, particularly children with special healthcare needs (CSHCN), who typically have higher and more complex clinical care requirements. CSHCN are at risk for chronic physical, developmental, behavioral, and/or emotional conditions and have a continuous need for healthcare services.[4] According to the Health Resources and Services Administration's report, more than 80% of CSHCN require prescription drugs in addition to the support for specialty care services, mental health services, occupational, physical, and speech therapies, and special medical equipment including aids for hearing, mobility, and communication.[5]

Providing adequate care to CSHCN requires frequent, accurate, and consistent health information capture and communications among all stakeholders including patients, caregivers, providers, home nurses, care coordinators, and more. Traditionally, transmitting out-of-clinic health information, such as presence and frequency of symptoms, adherence to medical therapies, and changes in clinical status, has occurred through phone triage calls, clinic appointments, or visits to the hospital. The fragmented nature of this communication may result in omissions or inconsistencies in the medical EHR notes. Reducing barriers to capturing health data outside of the clinic (ie, patient-generated health data [PGHD]) and automating information flow from the home environment into the EHR could improve care coordination, clinical decision-making, follow-up planning, and optimize health and patient/family quality of life, as outlined by patient, caregiver, and provider stakeholders in earlier research.[6,7] Furthermore, the COVID-19 pandemic has highlighted a greater need for and benefit of maintaining home-based care and monitoring,[8,9] though collecting relevant and complete health information at home remains challenging and requires an innovative approach.

Patient-generated health data

PGHD, such as medical diaries used by families to keep track of health information and events at home, has primarily been collected through mobile devices[10–12] and is used to support clinical decision-making[13] and improve quality of life.[14] In addition, communication and data-sharing technologies, such as patient portals,[15] cloud-based care plan sharing, and other approaches,[16,17] have been leveraged to enhance care coordination and management for children with chronic diseases. As mobile devices become more ubiquitous, care coordination and communications are increasingly supported by mobile health technologies.[18,19] PGHD is growing in quantity and usage, allowing for improved clinical decision-making and contributing to patient-reported outcomes within pediatric settings.[14,20–22] However, PGHD in the pediatrics population is infrequent and not publicly available, and its size could not be estimated from the published literature.

Structured PGHD (data that are stored in a predefined format, eg, home address, list of medication, sensor data) collected through the healthcare technologies can generally be integrated with EHR through currently established structure and standardized methods, such as, using Application programming interfaces (APIs) to transfer data into EHR flowsheets . In contrast, unstructured PGHD (eg, free-text notes, voice recordings) are expected to be reviewed and processed (put into a structure) individually by either patients, caregivers, or providers, thus introducing a significant burden of leveraging the data.[6] However, unstructured text may encapsulate rich information through narratives (contextual, semantic) which can be critical to understanding a patient's condition beyond what can be captured in a fixed structured format.[23]

Existing artificial intelligence (AI) technologies can be leveraged to overcome barriers in creating and processing unstructured PGHD. Unstructured text data can be generated through voice-interactive software, allowing patients and caregivers to easily generate notes with minimal effort or familiarity with data entry, a key

consideration for the overburdened families of chronically ill patients.[6] Furthermore, allowing narrative notes, whether spoken or typed, overcomes the limitations associated with predefined structured data entry, also enabling rich information capture. As expected, it takes an additional step to process the narrative notes, but natural language processing (NLP) and text mining approaches are potentially useful to digest and synthesize unstructured text.[23] The OurNotes project (part of OpenNotes initiative)[24] is a gateway for NLP-supported note processing to improve the extraction of valuable information from patient-generated notes and enhance shared clinical decision-making. In the literature, there is considerable promising exploration and application of NLP to clinical notes authored by clinicians and clinical staff to identify symptoms and conditions.[25–27] In contrast, NLP application to patient-generated notes has been limited.[28,29]

NLP literature

Automated analysis of free-text PGHD promises to ease documentation burdens, maximize value for patients, enable comprehensive patient information access for providers, and improve patient-clinician interaction. However, since PGHD often consists of entirely unstructured data with highly individualistic narration styles, the synthesis of PGHD to create meaningful and digestible information remains a challenge. A host of prior work has attempted to analyze free-text PGHD, using rule-based and deep-learning methods.[23] Rule-based methods are able to directly leverage existing work in the clinical domain, such as ontologies for accurate and relevant entity extraction, at the cost of generalizability over varied forms of free text.[26] Deep-learning solutions are able to account for variation in spelling and grammar by utilizing a probabilistic model of word/sentence meaning instead of deterministic rules[23] demonstrating better performance than rule-based methods.[30] Deep learning, however, typically requires costly data annotation and model training, leading to existing work on clinical information extraction (IE) being focused on narrow tasks such as drug event extraction. Some solutions have attempted to combine both deep-learning and rule-based methods but they do not focus on extracting the needed data for chronic disease management and still often require high-cost deep-learning model training.[23] For an IE pipeline to be helpful for chronic disease management using PGHD, it must extract symptom and drug information while being easily tuned to a specific cohort and be robust enough to capture various forms of text.

Objectives

In this paper, we introduce an IE pipeline which leverages a combination of rule-based methods and pretrained deep-learning NLP models to automatically extract information from caregiver or PGHD, specifically unstructured patient-reported medical notes collected at home. This hybrid pipeline is robust to misspellings and varied word representations while providing the ability to accommodate the needs of a specific patient, cohort, or clinician. We also propose a series of data integration methods that may aid in downstream data visualization, sharing, and assessment for caregiver/patient and provider.

METHODOLOGY—SMART SUMMARIZATION OF NOTES

Due to the informal nature of narrative notes, we chose to construct our IE pipeline around the scispaCy suite of deep-learning models which are pretrained on medical data with ontologies in mind.[31] The advanced scispaCy toolset considers representations of words according to their approximate meaning instead of performing a simple text match. This generalizability allows for rapid identification of key entities as well as attributes pertinent to these entities or to a sentence as a whole. Temporal information on health conditions is important for condition monitoring and clinical decision-making. Therefore, extracted note content is tagged with message timestamps to enable downstream visualizations in the form of easy-to-digest timelines of information key to communication for and care of patients with chronic diseases.[32] These "smart summary" timelines of the note content are available for 4 major information categories: drug, symptom, other qualitative entities (OQLEs), and other quantitative entities (OQNEs).

An overview of the proposed NLP pipeline is illustrated in Figure 1. This pipeline accepts user-generated data in the form of text or voice (upper green box). We then use an NLP model pretrained on clinical data[31] to extract key entities and link to ontologies if possible (blue boxes). The key entity extraction can also be augmented by entering manually defined entities for a given patient or cohort (lower green box). The process of entering manually defined entities acts as a real-time interactive query in that the visualization content will be immediately responsive to manually entered entities. Finally, we leverage the dependency parsing capabilities of the NLP model to extract further detail regarding each entity (yellow boxes) and the extracted data are aggregated and presented (white box).

The model

We employ the scispaCy package which contains spaCy models pretrained to process biomedical and clinical text.[31] SpaCy uses deep-learning methods, namely convolutional neural networks, to create generalized tools that can be applied to individual steps in an NLP pipeline.[33] We particularly leverage the named entity recognition (NER) and dependency parsing capabilities of scispaCy.

NER refers to the extraction and identification of key entities within a text span. ScispaCy reviews each term (word or phrase) in a given text span and predicts the chance that a given term is an entity. In cases where this prediction has high confidence, scispaCy further attempts to predict a Unified Medical Language System (UMLS) ID for the entity in a subprocess referred to as Named Entity Linking.[34] We also allow customization of the NER process through the addition of manually defined entities alongside those flagged by scispaCy, allowing the pipeline to be fit to the needs of a given cohort.

Dependency parsing refers to the prediction of the grammatical structure tree for a given text span. For each word in the text span, scispaCy assigns a link to either a child word or a parent word. These links create a tree-like structure that can be traced to identify subsentence components such as clauses (phrases), or, more directly, word-to-word dependencies. In addition, this task is trained alongside the subtask of part-of-speech tagging which labels words based on what grammatical form they most likely take (ie, Noun, Verb, Determinant). Figure 2 depicts one example of how dependency trees are used within our pipeline.

CASE STUDY: CYSTIC FIBROSIS

Cystic fibrosis cohort

We applied our pipeline to a simulated case of a child with cystic fibrosis (CF) to demonstrate its applicability and value. CF is a prevalent health condition among CSHCN given the need for frequent,

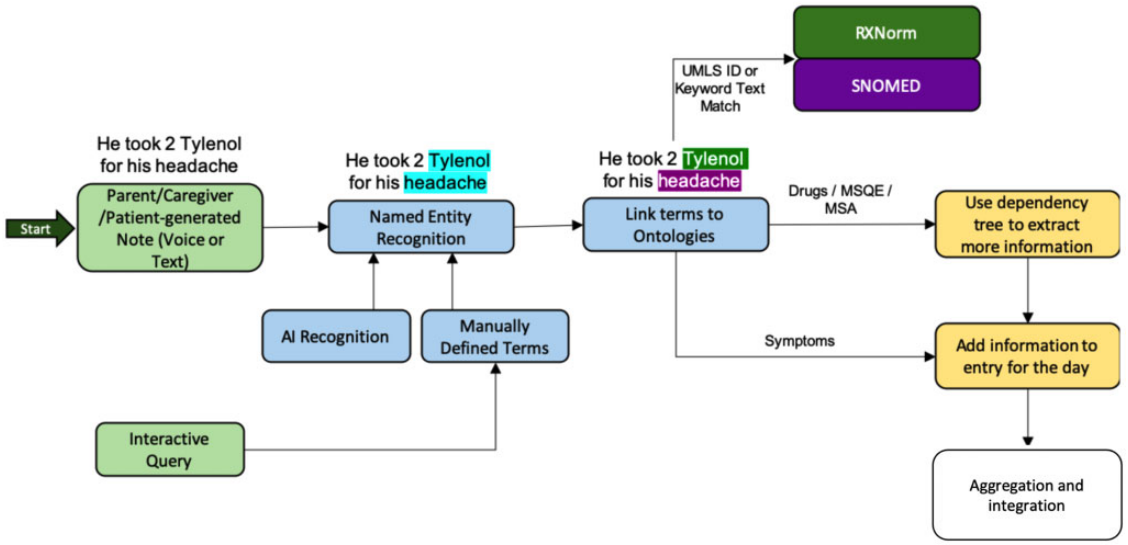

Figure 1. Process flow of note processing and information extraction.

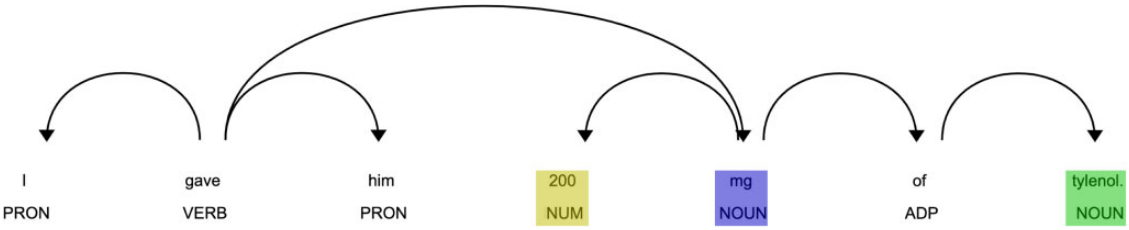

Figure 2. An example for drug dosage extraction using dependency trees. The NER model identifies the drug within the sentence, that is, Tylenol (green). Once identified, we move up the dependency tree until we either find an NUM or NOUN. If we find an NOUN (blue), we see if there is an NUM as a child to the NOUN. It is assumed that this NUM child (yellow) is the quantity and the NOUN (blue) is the unit for the drug's dosage. If no NOUN or NUM occurs in the same clause of the sentence, or subsection of the dependency tree, then no dosage information is extracted. NER: named entity recognition.

intensive care in the home environment. More than 30 000 people live with CF in the United States, with more than 70 000 patients worldwide. Approximately 1000 new cases of CF are diagnosed each year, with the vast majority of patients diagnosed before the age of 2.[35]

Daily care of an infant or child with CF generally requires the administration of frequent medications and chest therapies in order to optimize nutrition and lung health. Most children with CF require enzyme medications with every meal, along with supplemental vitamins and salt. Children with CF require increased caloric intake compared with their non-CF peers in order to maintain appropriate growth. Optimal lung health requires regular chest therapies involving several nebulized medications in addition to 20–30 min of chest physiotherapy (CPT) multiple times per day. During infancy, CPT is performed manually, requiring a significant amount of time and effort by caregivers.

To manage these children clinically, medical providers ask parents to monitor for signs of nutritional intake (amount, frequency of feeds), gastrointestinal malabsorption (frequency of bowel movements, presence of grease or oil in stools), and respiratory symptoms (cough, heavy breathing, wheezing). Depending on the age of the child and presence of symptoms, clinical visits may be as frequent as every 2 weeks or may stretch as long as every 3 months. Caregiver reports of the symptoms and findings noted above are vital to clinical decision-making though most often rely on memory or require a caregiver-initiated phone call or message to discuss concerns between medical visits. Accurate reporting can be particularly challenging when multiple people are responsible for the child's care at home (parents, grandparents, daycare, etc.). Practically, it is difficult to consistently and accurately document or provide health updates for CF patients in a timely manner to best address clinical concerns. Caregivers often have their hands full as they are administering medications, feeding their babies, and maintaining respiratory treatments. This suggests the need for a new approach to collect and communicate relevant health information to the medical team, which may allow for quicker clinical intervention when symptoms of malabsorption or respiratory concerns arise, potentially improving clinical outcomes and quality of life.

Note creation and IE overview

In order to evaluate our pipeline, we created notes which simulate the daily care for a young boy (Jon Doe) with CF. Each day has 2 notes, totaling 30 notes over a simulated 15-day period. We opted to use this synthetic dataset over true notes in order to target edge cases that may never arise in a true dataset and to lower data acquisition costs. Since this pipeline is oriented to being customizable for a given patient or cohort, the data collection process would necessitate the long-term commitment of a small subset of patient families. Instead, each note was constructed realistically with the support of a CF clinician (KK) to represent patterns that may be common for quick notation, such as text entry on a phone or verbal entry via

speech-to-text. In real terms, this patterning refers to shorter, occasionally incomplete sentences as well as the use of shorthand for various terminology. Likewise, we attempt to diversify the types of sentences to mimic the diversity of usage that may occur. Even so, it is important to recognize this diversity of patterns is not exhaustive. Management of CF involves nutritional and medicational considerations as well as the regular practice of therapies and tracking of symptoms. In the following sections, we describe how our pipeline extracts each of these components from our simulated PGHD.

In the following sections, we describe the types of information our pipeline extracts as well as how each category may lend itself to customization for a given cohort. While some of the entities in these categories are extracted automatically using NER and ontologies, all of the following categories allow the clinician or parent to manually define terms to highlight as entities and then process accordingly. These manually defined entities can be entered before or after notes are generated. This flexibility allows the IE pipeline to be tailored to fit a given cohort/patient situation as new questions arise. Table 1 provides an IE pipeline focusing on extraction of symptom, drug, OQLE, and OQNE categories.

Symptom extraction

One of the meaningful applications of this pipeline is for symptom extraction. Here, we take the gathered entities and compare them against the Systematized Nomenclature of Medicine (SNOMED) ontology,[26] specifically, the class names within the subtree of Clinical Findings.[36] For the entities with UMLS IDs, we directly compare them against the UMLS IDs of the SNOMED classes, allowing for an exact match. Otherwise, we compare the raw entity text against the raw class name texts to identify a match. If the given entity has a match within the SNOMED Clinical Findings subtree, it is marked as a symptom. Many of the symptoms of concern for a CF patient are described within the SNOMED ontology as standardized terms (eg, "fever" and "diarrhea") and were automatically identified via the NER lookup.

OQLE extraction

To further target ailments and nutritional considerations common in CF patients, we add OQLEs to allow the physician to get a quick snapshot of key terms (ie, tracking all references to "stool" to gather an understanding of the patient's bowel movements). To identify OQLEs, such as "Pediasure" for tracking nutritional information, we directly scan the text instead of focusing only on automatically extracted entities. Using the set of manually defined entities entered by the parent, caregiver, and/or physician, our algorithm searches for these terms within each note. Then, leveraging the dependency tree, we identify child or parent adjectives related to the key term. These adjectives serve as qualitative descriptors and are presented as modifiers to the given term. These descriptors are aggregated in the final presentation with included time-stamp information.

Drug/supplement and dosage extraction

To identify drug entities, we compare the extracted entities against RXNorm.[37] If there is a match, the entity is considered to be a drug. We then leverage the dependency tree to identify the relevant dosage information by moving up the tree from the drug entity to the first span of parent terms which are numbers. To identify the unit of measurement, we similarly move up the tree and identify the first noun. If no noun occurs before the first number, it is assumed that no unit of measurement was given.

Drug information encompasses a variety of units (eg, capsules or puffs) and can be mentioned both in terms of generic name as well as brand name. While our system was able to automatically identify generic names as drugs, some brand names or common shorthands (ie, "neb" for nebulizer) needed to be manually entered for tracking. Nutritional supplements, much like typical drugs, require the system to gather unit information. As such, we added manually defined entities to the system with the names of nutritional supplements being used by our hypothetical patient, allowing information to be extracted as it would for our drug dosage IE pipeline.

OQNE extraction

OQNE extraction uses a similar methodology to OQLE detection by collecting a user-defined set of OQNEs that are then searched for in each note. Once identified, we proceed to collect adjective information as well as quantity and unit information. This quantity information is extracted using a similar method to drug dosage extraction but allows broader forms of quantitative data, such as duration, frequency, or dosage. The flexibility of this algorithm to allow for user-specified input allows for customization specific to each disease, patient, and care institution. To target this cohort, we build a set of OQNEs that target CF treatments (ie, High Frequency Chest Wall Oscillation [Vest] therapies) which require both quantifiable dosage and qualitative attribute information.

Summary of extracted entities

Given these drugs, OQNEs, OQLEs, and symptoms, we were then able to generate and analyze a simulated set of parent- or patient-provided notes. Examples of terms in each of these categories, as well as an example sentence in which they are used, are shown in Table 2. As information regarding each of these 4 categories is extracted, we populate a series of tables that organize the information longitudinally. An example of the type of notes generated, as well as the form of the extracted information, is presented in Table 2.

Deployment and implementation

The proposed pipeline aims to extract meaningful information from patient-generated notes and present them in a digestible and actionable manner. The presentation could be through charts and/or visu-

Table 1. Information extraction pipeline focusing on extracting 4 categories of entities alongside quantitative and qualitative descriptors for each

Category	Type of information extracted	Entity recognition method
Symptom	Qualitative	NER and manually defined
Drug	Quantitative	NER and manually defined
Other qualitative entities (OQLEs)	Qualitative	Manually defined
Other quantitative entities (OQNEs)	Quantitative and qualitative	Manually defined

Note: These entities can be manually defined or detected automatically using NER linked to ontologies.
NER: named entity recognition.

Table 2. Examples of extracted entities, definitions, types, and sentences in each category

Extracted entities	Entity definition[a]	Related information	Type	Example sentences
Pediasure	Manual	–	Drug/supplement	"Took his Pediasure before bed time, and had 3 vitamins."
Vitamins	Manual	Quantity: "3"		
Vest	Manual	Quantity: "2"	OQNE	"Jon had the vest two times today."
Albuterol	Automatic	Dosage: "5 mg"	Drug	"Took his albuterol, pulmozyme, and tobi."
Pulmozyme	Automatic	–		
Tobramycin	Manual	–		"I gave him 5 mg of albuterol today."
Diaper	Manual	Diaper details: "wet"	OQLE	"Changed the wet diaper twice with loose stool before noon."
Stool	Manual	Stool details: "loose"		
Coughing	Automatic	Coughing details: "brief"	Symptom	"Jon had a brief coughing fit just now."

[a]Entity definitions are "Automatic" if they are present within the ontologies used (SNOMED and RXNORM). "Manual" refers to entities that were added manually via interactive customization to target this specific cohort and simulated patient.

OQLE: other qualitative entity; OQNE: other quantitative entity; SNOMED: Systematized Nomenclature of Medicine.

alizations (a mock-up data visualization tool is shared in Supplementary Appendix S1), which could improve symptom awareness and facilitate health communications and interpretability of the health data by patients, caregivers, and providers.[32] The ultimate goal of developing this pipeline is the adoption by all stakeholders in the care of CSHCN, including patients, caregivers, and providers. The algorithm could be deployable on patient or caregiver phones via a web portal, or more conveniently via an app, where users take medical notes and track symptoms.[6] Eventually, the integration of this pipeline within the EHR may ease the dependencies and burdens of patients and enable providers to access patient information and improve clinical care decisions in a timely manner.

The 21st Century Patients Cures Act allows access to personal health information and enables sharing with 3rd party apps through APIs.[38] Mobile app and PHGD integration with an EHR are increasingly being supported using interoperability standards (eg, Fast Healthcare Interoperability Resources [FHIR]).[39] However, from a health institution perspective, integrating external platforms with the EHR might be challenging due to interference with existing clinical workflow, internal security, and privacy measures and protocols which may vary for each health institution and EHR systems. As an alternative approach, patient portals could be leveraged for data sharing. Patient portal adoption has drastically increased due to the rapid deployment and adoption of telehealth during the COVID-19 pandemic. Therefore, apps to collect and manage notes generated by a patient could be linked to the patient's EHR, potentially through authentication of the patient portal (eg, Epic App Orchard API). Special considerations are needed for patients seeking care at multiple provider organizations.

DISCUSSIONS

Principal findings

The pipeline presented in this paper provides a viable avenue for patient-initiated health tracking that prioritizes ease-of-use for both the patient and clinician end-users. By leveraging contemporary NLP methodologies, we are able to process patient-generated text (eg, free-text notes, voice transcripts), create structured information, and organize the information for tabular and graphical presentation. Existing PGHD IE approaches that focus exclusively on deep-learning or rule-based methods miss out on rapid deployment and generalizability, respectively.[23,26] Hybrid approaches attempt to mitigate these issues yet still typically require costly deep-learning model training and have limited customization.[23] Likewise, few of these approaches directly target the critical use case of chronic disease management, and the need to comprehensively extract symptoms, drugs, treatments, and patient-specific information.[23] Unlike previous work, our approach focuses on extracting information relevant to chronic health conditions while allowing customization—manual entry of specific target terms—to maximize the utility of the tool.

In addition, since the pipeline leverages publicly available deep-learning models and ontologies, it can be easily replicated and customized for different cohorts and institutions. As the NLP model used has been pretrained on clinical data, replication of this pipeline has no cost for training and annotating data for the model, as well as limited need for deep-learning expertise. Likewise, the use of ontologies for entity linking allows the model to work with different clinical domains and ontologies, ensuring the entities identified are pertinent to the given cohort and task. We currently employ ontologies that target drugs and symptoms (RXNORM and SNOMED, respectively), but this pipeline can be expanded to include more specific ontologies (eg, NCIT for cancer; APAONTO for psychology).

From a practical standpoint, the use of the NLP pipeline may increase the amount of clinical data recorded as families of CSHCN are able to easily identify health events (eg, symptoms and medication changes) from the notes. Away from the clinic, such a processing mechanism could improve health management and eventually aid in adherence and early symptom detection. Iqbal et al[40] emphasized the value of remote care and alert systems being effective in reducing hospitalization. Clinically, integrating the proposed NLP pipeline with the EHR would allow providers to effectively observe the changes which are not easily and completely available from anecdotal notes, such as nutrition flowsheets in the EHR which depend on intermittent triaging to be completed. Using the integrated tools, care coordinators, nurses and clinicians would be able to access a more holistic view of a patient's health journey, improve care coordination and communications, and effectively intervene to improve clinical care.

Potential applications extend beyond the demonstrated use case of the CF and could be highly beneficial to any population impacted by chronic medical conditions. For any families with CSHCN, keeping track of health events out of the hospital is a necessary compo-

nent for patients and caregivers to maintain health communications with providers and clinical decision-making.[12,20,22,39]. Especially, if the patients are seeing multiple providers, required documentation and information during home care for each provider may potentially create burden, lead to recall bias, and adversely affect patient-provider communications.[41,42] With integration to mobile apps[19] or voice-interactive platforms,[6,7] daily note keeping burden can be substantially reduced. Furthermore, shared apps and platforms could be nested in a family's digital ecosystem where patients and caregivers can conveniently track health events. In addition, health management systems in sync with PGHD and consolidating outpatient communication may improve interpretation of free-text PGHD by the providers.[43] In that regard, integrated systems with apps and EHRs would support the shared decision-making and align with the OpenNotes initiative to improve patient healthcare communications.[24] The pipeline has the potential to be used by health systems to support remote care, symptom tracking, and adherence, which also fits with the Creating Opportunities Now for Necessary and Effective Care Technologies (CONNECT) for Health Act.

Technical limitations and special considerations

While this pipeline allows customization via the addition of manual entities, the functionality over these entities (manual or automatic) cannot be customized. For example, while the considered set of drug terms can be expanded, this algorithm is limited regarding where it can subsequently acquire dosage and unit information. One key limitation is the fact that the current methodology does not support entailment between sentences. In the example "I gave him midazolam. He took 1 pill," the algorithm would be unable to identify "1 pill" as a reference to "midazolam" due to the reference being in a separate sentence, and thus a different dependency tree. Likewise, unexpected grammatical structure can additionally cause erroneous values. To address this in future iterations of the pipeline, we could improve the error mitigation by creating simple functionality to "ignore future results like these" for end-users to mark in order to filter out similar results based on the underlying structure.

Moreover, our current evaluation metrics are limited in their scope due to each version of the system being modified for a given set of patients. Namely, we iteratively add entities to fit information that was originally missed by the model. This, in turn, creates a situation where a given evaluation can either perform better through more thorough keyword augmentation or perform worse in a more generalized scenario. In addition, our limited set of 30 synthetic notes bottlenecks the extent of our evaluations. In the future studies, we plan to collect real-world data and test our pipeline to produce generalizable results.

CONCLUSION

With recent advances in medically oriented, pretrained, and publicly available NLP models, such as scispaCy,[31] it is possible to parse sentences according to their grammatical structure and identify terms present in given ontologies. This allows the creation of structured data from otherwise unstructured notes, with no requirement to identify targeted entities and attributes in advance. The data can be integrated with the EHR and visualized for patients, caregivers, and providers to track and manage healthcare activities. Altogether, the proposed pipeline can lower the burden for remote care and chronic disease management associated with CSHCN and improve utilization of unstructured PGHD.

FUNDING

This work was partially supported by the Health Resources and Services Administration Maternal and Child Health Bureau Grand Challenge for Care Coordination for CSHCN (grant Number 720467022000) and CTSA (grant number RUL1TR02733).

AUTHOR CONTRIBUTIONS

ES and YH conceived the presented idea. All authors contributed to the design of the work. ES, S-AH, and KK contributed to the acquisition and interpretation of the data. S-AH developed and tested the model. S-AH and ES drafted the manuscript. Collectively, S-AH, ES, KK, JL, SR, and YH critically reviewed and revised the manuscript. The final version of the manuscript is approved by the authors.

SUPPLEMENTARY MATERIAL

Supplementary material is available at *JAMIA Open* online.

CONFLICT OF INTEREST STATEMENT

None declared.

DATA AVAILABILITY

The data underlying this article will be shared on reasonable request to the corresponding author.

REFERENCES

1. Lau HS, Florax C, Porsius AJ, De Boer A. The completeness of medication histories in hospital medical records of patients admitted to general internal medicine wards. *Br J Clin Pharmacol* 2000; 49 (6): 597–603.
2. Bell SK, Delbanco T, Elmore JG, *et al.* Frequency and types of patient-reported errors in electronic health record ambulatory care notes. *JAMA Netw Open* 2020; 3 (6): e205867.
3. Khoo EM, Lee WK, Sararaks S, *et al.* Medical errors in primary care clinics—a cross sectional study. *BMC Fam Pract* 2012; 13 (1): 127.
4. McPherson M, Arango P, Fox H, *et al.* A new definition of children with special health care needs. *Pediatrics* 1998; 102 (1 Pt 1): 137–40.
5. 2009–C. The National Survey of Children with Special Health Care Needs. https://mchb.hrsa.gov/sites/default/files/mchb/Data/NSCH/nscshcn0910-chartbook-jun2013.pdf Accessed June 9, 2021.
6. Sezgin E, Noritz G, Lin S, Huang Y. Feasibility of a voice-enabled medical diary app (SpeakHealth) for caregivers of children with special health care needs and health care providers: mixed methods study. *JMIR Form Res* 2021; 5 (5): e25503.
7. Sezgin E, Noritz G, Elek A, *et al.* Capturing at-home health and care information for children with medical complexity using voice interactive technologies: multi-stakeholder viewpoint. *J Med Internet Res* 2020; 22 (2): e14202.
8. Watson AR, Wah R, Thamman R. The value of remote monitoring for the COVID-19 pandemic. *Telemed J E Health* 2020; 26 (9): 1110–2.
9. Sezgin E, Huang Y, Ramtekkar U, Lin S. Readiness for voice assistants to support healthcare delivery during a health crisis and pandemic. *NPJ Digit Med* 2020; 3: 122.
10. Park YR, Lee Y, Kim JY, *et al.* Managing patient-generated health data through mobile personal health records: analysis of usage data. *JMIR Mhealth Uhealth* 2018; 6 (4): e89.
11. Chan Y-FY, Bot BM, Zweig M, *et al.* The asthma mobile health study, smartphone data collected using ResearchKit. *Sci Data* 2018; 5: 180096.

12. Chew C-C, Hss A-S, Chan H-K, Hassali MA. Medication safety at home: a qualitative study on caregivers of chronically ill children in Malaysia. *Hosp Pharm* 2020; 55 (6): 405–11.
13. Jim HSL, Hoogland AI, Brownstein NC, *et al.* Innovations in research and clinical care using patient-generated health data. *CA Cancer J Clin* 2020; 70 (3): 182–99.
14. Petersen C. Patient-generated health data: a pathway to enhanced long-term cancer survivorship. *J Am Med Inform Assoc* 2016; 23 (3): 456–61.
15. Fiks AG, DuRivage N, Mayne SL, *et al.* Adoption of a portal for the primary care management of pediatric asthma: a mixed-methods implementation study. *J Med Internet Res* 2016; 18 (6): e172.
16. Desai AD, Wang G, Wignall J, *et al.* User-centered design of a longitudinal care plan for children with medical complexity. *J Am Med Inform Assoc* 2020; 27 (12): 1860–70.
17. Desai AD, Jacob-Files EA, Wignall J, *et al.* Caregiver and health care provider perspectives on cloud-based shared care plans for children with medical complexity. *Hosp Pediatr* 2018; 8 (7): 394–403.
18. Tiase VL, Sward KA, Del Fiol G, Staes C, Weir C, Cummins MR. Patient-generated health data in pediatric asthma: exploratory study of providers' information needs. *JMIR Pediatr Parent* 2021; 4 (1): e25413.
19. Baysari MT, Westbrook JI. Mobile applications for patient-centered care coordination: a review of human factors methods applied to their design, development, and evaluation. *Yearb Med Inform* 2015; 10 (1): 47–54.
20. Miller MW, Ross RK, Voight C, *et al.* Patient-generated digital images after pediatric ambulatory surgery. *Appl Clin Inform* 2016; 7 (3): 646–52.
21. Coons SJ, Eremenco S, Lundy JJ, O'Donohoe P, O'Gorman H, Malizia W. Capturing patient-reported outcome (PRO) data electronically: the past, present, and promise of ePRO measurement in clinical trials. *Patient* 2015; 8 (4): 301–9.
22. Patient-Generated Health Data. https://www.healthit.gov/topic/scientific-initiatives/pcor/patient-generated-health-data-pghd Accessed August 23, 2021.
23. Dreisbach C, Koleck TA, Bourne PE, Bakken S. A systematic review of natural language processing and text mining of symptoms from electronic patient-authored text data. *Int J Med Inform* 2019; 125: 37–46.
24. OurNotes: Patients and Clinicians Creating Notes Together. https://www.opennotes.org/ournotes/ Accessed June 9, 2021.
25. Topaz M, Radhakrishnan K, Blackley S, Lei V, Lai K, Zhou L. Studying associations between heart failure self-management and rehospitalizations using natural language processing. *West J Nurs Res* 2017; 39 (1): 147–65.
26. Gaudet-Blavignac C, Foufi V, Bjelogrlic M, Lovis C. Use of the Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) for processing free text in health care: systematic scoping review. *J Med Internet Res* 2021; 23 (1): e24594.
27. Pivovarov R, Elhadad N. Automated methods for the summarization of electronic health records. *J Am Med Inform Assoc* 2015; 22 (5): 938–47.
28. Iatraki G, Kondylakis H, Koumakis L, *et al.* Personal Health Information Recommender: implementing a tool for the empowerment of cancer patients. *Ecancermedicalscience* 2018; 12: 851.
29. McNeer E, Beck C, Weeks HL, *et al.* Building longitudinal medication dose data using medication information extracted from clinical notes in electronic health records. *J Am Med Inform Assoc* 2021; 28 (4): 782–90.
30. Wei Q, Ji Z, Li Z, *et al.* A study of deep learning approaches for medication and adverse drug event extraction from clinical text. *J Am Med Inform Assoc* 2020; 27 (1): 13–21.
31. Neumann M, King D, Beltagy I, Ammar W. ScispaCy: fast and robust models for biomedical natural language processing. In: Proceedings of the 18th BioNLP Workshop and Shared Task. Association for Computational Linguistics; 2019: 319–27. doi: 10.18653/v1/w19-5034. https://arxiv.org/abs/1902.07669.
32. Vaughn J, Kamkhoad D, Shaw RJ, Docherty SL, Subramaniam AP, Shah N. Seriously ill pediatric patient, parent, and clinician perspectives on visualizing symptom data. *J Am Med Inform Assoc* 2021; 28 (7): 1518–25.
33. Honnibal M, Montani I. spaCy 2: natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. https://sentometrics-research.com/publication/72/ Accessed August 23, 2021.
34. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res* 2004; 32 (Database issue): D267–70.
35. 2019-Patient-Registry-Annual-Data-Report.pdf. https://www.cff.org/Research/Researcher-Resources/Patient-Registry/2019-Patient-Registry-Annual-Data-Report.pdf Accessed August 23, 2021.
36. Schulz S, Klein GO. SNOMED CT—advances in concept mapping, retrieval, and ontological foundations. Selected contributions to the Semantic Mining Conference on SNOMED CT (SMCS 2006. *BMC Med Inform Decis Mak* 2008; 8 (Suppl 1): S1.
37. Nelson SJ, Zeng K, Kilbourne J, Powell T, Moore R. Normalized names for clinical drugs: RxNorm at 6 years. *J Am Med Inform Assoc* 2011; 18 (4): 441–8.
38. Bonamici S. 21st Century Cures Act. 2016. https://www.congress.gov/bill/114th-congress/house-bill/34 Accessed June 24, 2021.
39. Tiase VL, Hull W, McFarland MM, *et al.* Patient-generated health data and electronic health record integration: a scoping review. *JAMIA Open* 2020; 3 (4): 619–27.
40. Iqbal FM, Lam K, Joshi M, Khan S, Ashrafian H, Darzi A. Clinical outcomes of digital sensor alerting systems in remote monitoring: a systematic review and meta-analysis. *NPJ Digit Med* 2021; 4 (1): 7.
41. Ranade-Kharkar P, Weir C, Norlin C, *et al.* Information needs of physicians, care coordinators, and families to support care coordination of children and youth with special health care needs (CYSHCN). *J Am Med Inform Assoc* 2017; 24 (5): 933–41.
42. Glick AF, Farkas JS, Nicholson J, *et al.* Parental management of discharge instructions: a systematic review. *Pediatrics* 2017; 140 (2): e20164165.
43. Ye J. The impact of electronic health record–integrated patient-generated health data on clinician burnout. *J Am Med Inform Assoc* 2021; 28 (5): 1051–6.