International Journal of Medical Informatics 177 (2023) 105144

Contents lists available at ScienceDirect

International Journal of Medical Informatics

journal homepage: www.elsevier.com/locate/ijmedinf

ReDWINE: A clinical datamart with text analytical capabilities to facilitate rehabilitation research

David Oniani[a], Bambang Parmanto[a], Andi Saptono[a], Allyn Bove[c], Janet Freburger[c], Shyam Visweswaran[d,e,f], Nickie Cappella[d,f], Brian McLay[d,f], Jonathan C. Silverstein[d,f], Michael J. Becich[d,f], Anthony Delitto[c], Elizabeth Skidmore[b], Yanshan Wang[a,d,e,f,*]

[a] *Department of Health Information Management, University of Pittsburgh, Pittsburgh, PA, USA*
[b] *Department of Occupational Therapy, University of Pittsburgh, Pittsburgh, PA, USA*
[c] *Department of Physical Therapy, University of Pittsburgh, Pittsburgh, PA, USA*
[d] *Department of Biomedical Informatics, University of Pittsburgh School of Medicine, Pittsburgh, PA, USA*
[e] *Intelligent Systems Program, University of Pittsburgh, Pittsburgh, PA, USA*
[f] *Clinical and Translational Science Institute, University of Pittsburgh, Pittsburgh, PA, USA*

ARTICLE INFO

Keywords:
Electronic health records
Rehabilitation
Clinical research informatics
Informatics infrastructure
Natural language processing
Common data model

ABSTRACT

Rehabilitation research focuses on determining the components of a treatment intervention, the mechanism of how these components lead to recovery and rehabilitation, and ultimately the optimal intervention strategies to maximize patients' physical, psychologic, and social functioning. Traditional randomized clinical trials that study and establish new interventions face challenges, such as high cost and time commitment. Observational studies that use existing clinical data to observe the effect of an intervention have shown several advantages over RCTs. Electronic Health Records (EHRs) have become an increasingly important resource for conducting observational studies. To support these studies, we developed a clinical research datamart, called ReDWINE (Rehabilitation Datamart With Informatics iNfrastructure for rEsearch), that transforms the rehabilitation-related EHR data collected from the UPMC health care system to the Observational Health Data Sciences and Informatics (OHDSI) Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) to facilitate rehabilitation research. The standardized EHR data stored in ReDWINE will further reduce the time and effort required by investigators to pool, harmonize, clean, and analyze data from multiple sources, leading to more robust and comprehensive research findings. ReDWINE also includes deployment of data visualization and data analytics tools to facilitate cohort definition and clinical data analysis. These include among others the Open Health Natural Language Processing (OHNLP) toolkit, a high-throughput NLP pipeline, to provide text analytical capabilities at scale in ReDWINE. Using this comprehensive representation of patient data in ReDWINE for rehabilitation research will facilitate real-world evidence for health interventions and outcomes.

1. Introduction

According to the World Health Organization (WHO), rehabilitation is defined as a set of interventions designed to optimize functioning and reduce disability in individuals with acute and chronic diseases and conditions (e.g., stroke, chronic pain, surgery, cancer) as they interact with their environment" [1]. Researchers in rehabilitation have used randomized clinical trials (RCTs), the gold standard for evidence-based medicine, to study mechanisms of change and to provide evidence for the efficacy of rehabilitation interventions. However, the use of RCTs in rehabilitation research presents several challenges, such as lack of generalizability of the findings, high cost, and high time commitment [2].

Observational studies have been used to observe the effects of interventions in real-world settings without experimental assignment. Observational studies have several advantages over RCTs, including greater external validity and being faster and less expensive to conduct. With the emergence of electronic health records (EHRs) in major health care systems, EHRs have become an increasingly important resource for conducting observational studies [3]. Having access to EHR data from

* Corresponding author.
E-mail address: yanshan.wang@pitt.edu (Y. Wang).

https://doi.org/10.1016/j.ijmedinf.2023.105144
Received 28 March 2023; Received in revised form 14 June 2023; Accepted 6 July 2023
Available online 11 July 2023

large health care systems is particularly useful for rehabilitation research since many patients who are candidates for rehabilitation receive it across multiple settings. For example, an individual admitted to the hospital for a stroke may start rehabilitation in the acute care hospital but then may be transferred to an inpatient rehabilitation facility or skilled nursing facility to continue rehabilitation, followed by a transfer to the community where s/he may receive rehabilitation in the home, the outpatient setting, or both. Accessing EHR data from these various settings would provide a complete picture of the patient's rehabilitation journey and provide a more accurate picture of the value and effectiveness of rehabilitation.

To facilitate EHR-based observational studies, the Observational Health Data Sciences and Informatics (OHDSI) community has developed open-source standardization vocabulary and software tools, such as the Observational Medical Outcomes Partnership Common Data Model (OMOP CDM). The OMOP CDM has established a standardized method for mapping and integrating EHR data, as well as many other types of data (e.g., claims data, transactional data), into a universal format of data linking patients, measurements, observations, visits, and procedures. In addition, the OHDSI community also offers open-source data analytics tools to facilitate observational research. ATLAS [4] is a web-based tool that provides a user interface to visualize data and standardized vocabulary and to define cohorts on the observational data in the OMOP CDM format. HADES [4] is a data analytics tool that provides a set of R packages for large-scale analytics, including population characterization, population-level causal effect estimation, and patient-level prediction.

Unstructured narratives in the EHR, such as clinical notes and procedure notes, are critical in rehabilitation research. Notes provide a wealth of information about patients' medical histories, symptoms, treatments, and lifestyle factors that structured EHRs alone cannot capture [5]. Some estimates on how much valuable research data is contained in clinical notes range from 70 to 80 percent or higher [6]. For example, in a physical therapy session, exercise-specific information is only available in the clinical note [7]. Such information is valuable for clinical researchers who are interested in studying the association of different exercises and therapy outcomes. By automating the process of extracting information from unstructured EHRs, clinical and translational researchers can save time and increase their efficiency, while also gaining a more comprehensive view of patient data. The OMOP CDM has incorporated the representations associated with unstructured EHRs by introducing NOTE and NOTE_NLP tables, which store clinical textual data and the output from clinical NLP tools, respectively. However, the OHDSI community currently does not provide clinical NLP tools to extract medical concepts from unstructured EHRs.

In this article, we describe the development of a clinical research datamart prototype, called ReDWINE (Rehabilitation Datamart With Informatics iNfrastructure for rEsearch), that transforms rehabilitation-related EHR data collected in a large health care system to the OMOP CDM to facilitate rehabilitation research. The standardized EHR data stored in RedWINE could reduce the time and effort required to harmonize and map data and enable researchers to pool and analyze data from multiple sources, leading to more robust and comprehensive research findings. We also deployed the ATLAS and HADES tools in ReDWINE to facilitate cohort definition and clinical data analysis. In addition, we adopted and deployed the Open Health Natural Language Processing (OHNLP) toolkit [8], a high-throughput NLP pipeline, to provide text analytical capabilities at scale. Using this comprehensive representation of patient data in ReDWINE for rehabilitation research could facilitate real-world evidence for health interventions and outcomes.

2. Methods

In this section, we describe the implementation details of ReDWINE. ReDWINE is designed as a self-service informatics and data analytics platform that leverages EHR data from the University of Pittsburgh Medical Center (UPMC) and supports observational rehabilitation research at the School of Health and Rehabilitation Sciences (SHRS) at the University of Pittsburgh. Fig. 1 shows the overall ReDWINE architecture and data pipeline.

2.1. Data collection

The University of Pittsburgh School of Medicine's Department of Biomedical Informatics designed and implemented a research data warehouse called Neptune [9] that extracts, transforms and stores EHR data from UPMC as part of their Business Associates Agreement (BAA) with the UPMC. Neptune [9] has a data layer that contains structured data de-identified to HIPAA Limited Data with dates and unstructured EHR clinical text data for research purposes. Due to the complex nature of multiple UPMC EHR systems, EHR data are stored in Neptune with minimal filtering and processing. The ReDWINE team worked with the Neptune team and retrieved rehabilitation-related EHR data from the data layer and transformed them into the OMOP CDM for rehabilitation researchers.

As the first step in developing the ReDWINE prototype, we used EHR data from a cohort of patients diagnosed with stroke. The reason we selected stroke is that it is a prevalent disease and many rehabilitation researchers at SHRS focus on rehabilitation therapies post-stroke. We defined a cohort of patients diagnosed with stroke and admitted to a UPMC acute care hospital (see the supplemental file for ICD-10 codes) between January 1, 2016 and December 31, 2016 at UPMC. We collected any available EHR data on these patients for the 12 months following the stroke across the outpatient settings, including demographics, diagnosis billing codes, procedures codes, encounter notes, and procedure notes, which were created between January 1, 2016 and December 31, 2018. The University of Pittsburgh's Institutional Review Board (IRB) reviewed and approved development of ReDWINE (IRB #21040204).

2.2. Database

We used the PostgreSQL[1] database in our implementation for RedWINE since it is well-supported by the OHDSI community. PostgreSQL is a free and open-source object-relational database management system (ORDBMS) focused on extensibility and SQL compliance. In the database, we defined three schemas, namely RAW, CDM, and RESULTS. Every schema has a specific purpose and contains corresponding tables:

- RAW schema: As the name suggests, the RAW schema holds the raw data obtained from Neptune that is subsequently transformed, standardized, and mapped via the Extract, Transform, Load (ETL) process[2]. This schema is based on the EHR data received from Neptune and thus can be flexible in its design. The tables in this schema depend on the nature and modality of the data.
- CDM schema: The CDM schema contains standardized EHR data in OMOP format following the standardized vocabulary and tables defined in the OMOP CDM.
- RESULTS schema: The RESULTS schema contains information and statistics generated by the Automated Characterization of Health Information at Large-scale Longitudinal Evidence Systems (ACHILLES) tool[3], which validates data quality and determines whether data meet a given requirement.

Detailed implementation information about each schema in ReDWINE and the ETL process can be accessed in the supplementary file.

[1] https://www.postgresql.org/.

[2] https://ohdsi.github.io/TheBookOfOhdsi/ExtractTransformLoad.html.

[3] https://github.com/OHDSI/Achilles.

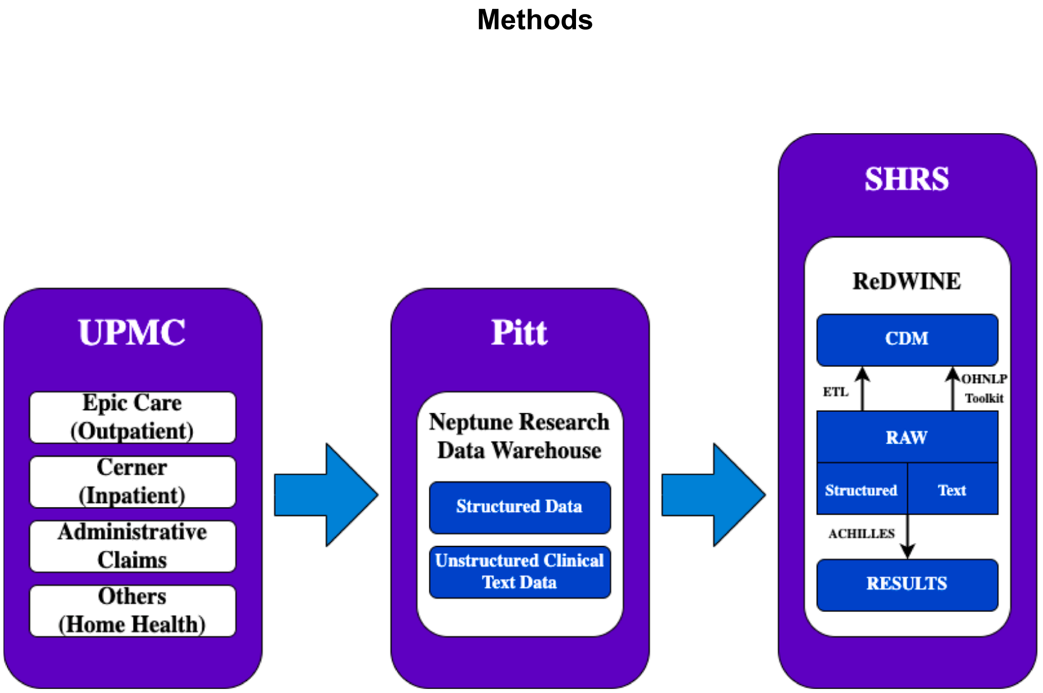


Fig. 1. ReDWINE Architecture and Data Pipeline.

Additionally, ReDWINE possesses text analytical capabilities that utilize high-throughput NLP pipelines to convert unstructured clinical texts into structured and standardized data, which are stored in the Note_NLP table. Details about the text analytical capabilities can be found in the supplemental file due to the word limit of the manuscript.

2.3. *Data quality assurance*

Because the data transformed between different standards and codes, we conducted the data quality assurance (QA) analysis to assess the quality of the ETL scripts and to ensure that there is no significant data loss in the pipeline. This is done by computing record counts of the original data in the RAW schema and in the CDM schema followed by generating statistics to assess data loss. The step ensures that the pipeline is robust and can handle future streams of data well.

3. Results

3.1. *Statistics of ReDWINE prototype*

Table 1 lists the statistics of the ReDWINE prototype using the stroke cohort. Currently, there is a total of 13,604 patients in the database, with 6,673 males and 6,931 females. There are ~ 2 million diagnoses entries for the cohort, with an average of 140 unique diagnoses per person. ReDWINE contains ~ 2 million clinical notes for the cohort and a total of ~ 304 million medical concepts after the NLP processing, with an average of 5,018 medical concepts per patient.

3.2. *Data QA results*

Table 2 shows the results of data QA analysis that assesses the quality of the ETL process and to make sure that there is no significant data loss. We found that the data transformation ETL scripts delivered good performance for demographics, diagnosis, encounter visits, and procedures with minimal data loss (ranges from 0% to 9.627%). For the RAW.demographics vs CDM.person comparison, there was no data loss and every single person was mapped from the table in the RAW schema to the corresponding table in the CDM schema. As for RAW.diagnoses vs CDM.condition_occurrence, there were 2,475 records left unmapped (0.129%). For RAW.encounters vs CDM.visit_occurence, all of the records were successfully mapped to the CDM table without data loss. Finally, for RAW.procedures vs CDM.procecure_occurrence, we were not

Table 1
Cohort statistics of the ReDWINE prototype.

	Number	Percentage
Gender	13,604	
Male	6,673	49%
Female	6,931	51%
Age		
18–34 years old	178	1%
35–44 years old	306	2%
45–54 years old	666	5%
55–64 years old	1,843	14%
65–74 years old	3,096	23%
75–84 years old	3,393	25%
>= 85 years old	4,016	30%
Race		
American Indian/Alaskan Native	14	0%
Asian	99	1%
Black or African American	1,325	10%
Native Hawaii/Pacific Islander	3	0%
White	11,661	85%
Unable to Provide	499	4%
Ethnicity		
Hispanic or Latino	64	0%
Not Hispanic or Latino	12,471	92%
Unable to Provide	1,069	8%
Diagnoses	1,832,139	–
Unique diagnoses per patient (average)	51	–
Visits	1,987,791	–
Visits per person (average)	146	–
Procedures	1,214,388	–
Procedures per person (average)	89	–
Notes	2,096,415	–
Notes per person (average)	154	–
Clinical NLP Concepts (Note_NLP table)	304,228,932 (60,627 unique)	–
Concepts per person (average)	5,018	–

Table 2
Results of data quality assurance analysis for ReDWINE prototype.

Comparison	RAW	CDM	Difference	RAW Lost (%)
RAW.demographics vs CDM.person	13,604	13,604	0	0
RAW.diagnoses vs CDM. condition_occurrence	1,911,650	1,909,175	2,475	0.129
RAW.encounters vs CDM. visit_occurrence	1,987,791	1,987,791	0	0
RAW.procedures vs CDM. procedure_occurrence	1,343,759	1,214,388	129,371	9.627

able to map 129,371 records since a number of CPT-4 codes that were not present in the OMOP CDM standardized vocabulary had to be excluded. This has resulted in the exclusion of 9.627% of the original 1,343,759 records. The reason that the data quality control procedure yields satisfactory results is that we adopted automated mapping and manual mapping approaches to handle different types of data elements during the ETL process. Further detailed information regarding the ETL process can be found in the supplement. Overall, the record counts are consistent between the RAW and CDM schemas, which indicates the data were successfully transferred to the OMOP CDM.

3.3. Case studies

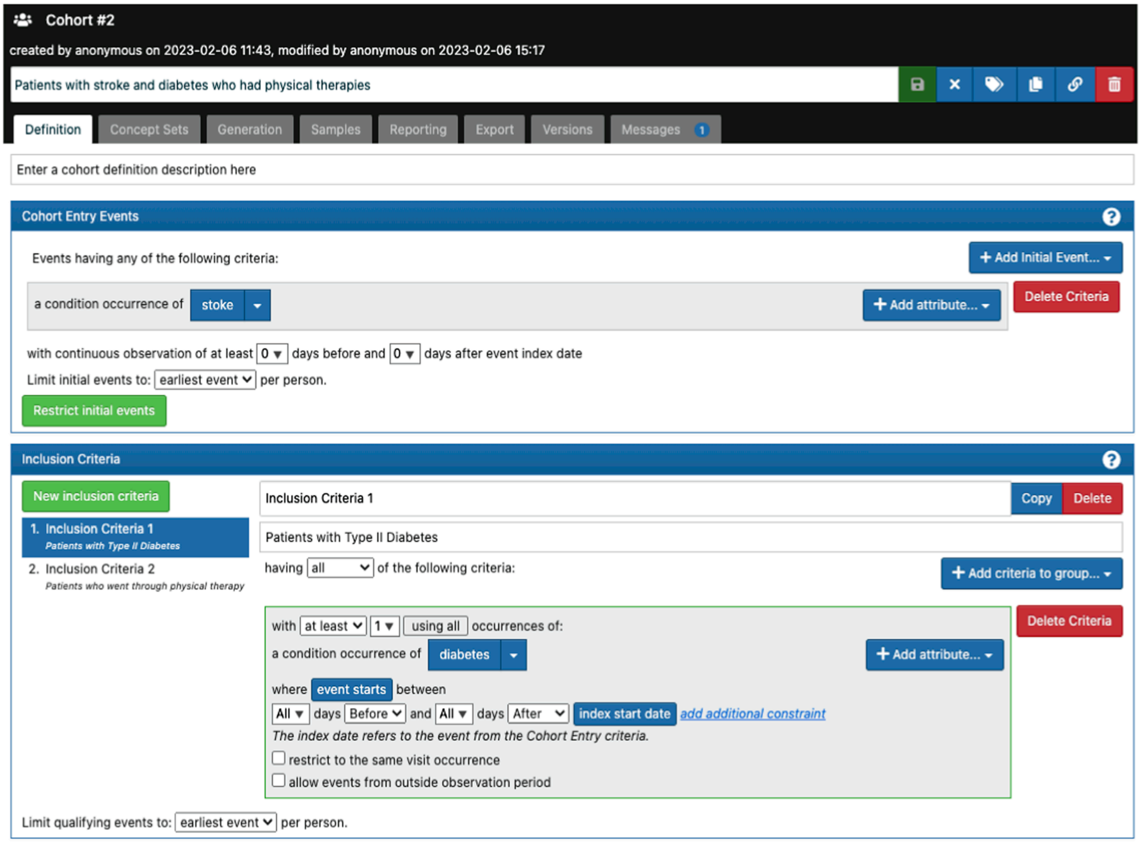
This section presents three case studies that showcase the utilization of ReDWINE for facilitating rehabilitation research. In Case Study 1, we demonstrate how a rehabilitation researcher can leverage ReDWINE's cohort discovery tools to define patient cohorts and idenfity patients based on the cohort definition. Case Study 2 illustrates the utilization of the general medical concepts extracted by ReDWINE's text analytical tool from clinical notes, which may not be available in the structured data, for patient identification. Furthermore, Case Study 3 showcases the deployment of a specific customized NLP algorithm within ReDWINE to efficiently extract customized concepts from millions of clinical notes. These case studies collectively highlight the potential of ReDWINE in supporting rehabilitation research and beyond.

3.4. Case study 1: Use ATLAS to define cohorts and visualize results

Suppose a rehabilitation researcher is interested in identifying a cohort of patients who had physical therapies and were diagnosed with type 2 diabetes followed by stroke, he/she could define these criteria using the cohort definition tool in ATLAS. The researcher could use the OHDSI Athena tool to identify the standardized OMOP CDM concept IDs to define three concept sets for the inclusion criteria. Using the defined concept sets, the researcher could define the cohort in ATLAS as shown in Fig. 2. The cohort entry event is defined as the cohort index time when subjects in the database are entered into the cohort. In our case, we used the stroke diagnosis (concept ID: 35207821) as the entry event. Then the researcher could define inclusion criteria to further restrict the cohort. In this case study, we used physical therapy (concept ID: 2314284) and type 2 diabetes (concept ID: 35206882). One could add more related concept IDs in each concept set, though we only used one concept ID for demonstration purposes. Then the researcher could generate the cohort in the Generation function. In this case study, we could identify five patients who met the cohort criteria.

3.5. Case study 2: Use NLP-extracted concepts to identify patients

Clinical information within unstructured EHRs can be very important for clinical and translational researchers. The NLP-extracted concepts in the ReDWINE could promote the use of such information for


Cohort #2
created by anonymous on 2023-02-06 11:43, modified by anonymous on 2023-02-06 15:17
Patients with stroke and diabetes who had physical therapies

Enter a cohort definition description here

Cohort Entry Events
Events having any of the following criteria:
a condition occurrence of stoke
with continuous observation of at least 0 days before and 0 days after event index date
Limit initial events to: earliest event per person.

Inclusion Criteria
1. Inclusion Criteria 1 — Patients with Type II Diabetes
2. Inclusion Criteria 2 — Patients who went through physical therapy

Inclusion Criteria 1
Patients with Type II Diabetes
having all of the following criteria:
with at least 1 using all occurrences of:
a condition occurrence of diabetes
where event starts between All days Before and All days After index start date add additional constraint
The index date refers to the event from the Cohort Entry criteria.
□ restrict to the same visit occurrence
□ allow events from outside observation period
Limit qualifying events to: earliest event per person.

Fig. 2. Cohort definition inside Atlas.

observational studies. In a case study, we assume that a researcher wanted to identify patients who experienced one or more falls. We used the Note_NLP table and the note_nlp_concept_id “436583” to identify the patients who had fall events mentioned in the clinical notes. We identified 10,322 patients using the HADES tool. We loaded these patients’ data into an R dataframe for further analysis. We also found that in the Note_NLP table there were 358 unique lexical variants of the fall event as documented in clinical notes. Examples of different variants include “accident &fell”, “Accident a fall”, “accident and fall”, “accident and falling”, “accident and fell”, “accident from falling”, “accident/ fall”, “accidents/falls”, “down Falls”, “down Fall”, “down and falling”, “down and falls”, “fall”, and “fall and injured”. The R script to identify these patients can be found in the supplemental file.

3.6. Case study 3: Running a specific customized NLP algorithm in ReDWINE

In Case Study 2, the NLP-extracted concepts are derived from a general medical ontology (i.e., OMOP Standardized Vocabulary). However, it is important to note that these general medical concepts may not always meet the specific requirements of clinical studies due to their limitations, such as a lack of granularity. Therefore, the ability to run a specific customized NLP algorithm again all clinical notes in the database is critical for clinical research. In a previous study [7], we developed a customized rule-based NLP algorithm to extract physical rehabilitation exercise information from clinical notes. The physical rehabilitation exercise information included 101 medical concepts across nine categories, including type of motion, side of body, location on body, plane of motion, duration, set and rep information, exercise purpose, exercise type, and body position. We deployed this customized NLP algorithm in ReDWINE and ran it again all clinical documents in ReDWINE. Finally, it took ∼ 16 h to run the algorithm on the ∼ 2 m clinical notes. All the NLP-extracted concepts were populated to the Note_NLP table. This case study demonstrates how ReDWINE harnesses its text analytical capabilities to efficiently and effectively run customized NLP algorithms. These capabilities pave the way for conducting large-scale studies in the field of rehabilitation research [10].

4. Discussion

Given the availability of increasing volumes of EHR data and improving computing power, the ability to access and utilize near real-time data for both research and clinical purposes has become a reality. We present an approach that can be used to harmonize and standardize EHR data from multiple settings and that uses different EHR software (e.g., Epic, Cerner) using rehabilitation care for stroke as a use case example. Precision rehabilitation has recently gained interest in both research and practice that seeks to deliver the precise and individualized intervention at the right time [11]. The information technology infrastructure, standardized data elements, common data model, and data analytical platforms developed in ReDWINE and observational studies enabled by ReDWINE could be leveraged to advance precision rehabilitation research and ultimately to maximize function and minimize disability in patients with acute illness or injury and chronic conditions. Furthermore, ReDWINE offers a platform for learning health systems research that seeks to link health system data and experiences with external evidence to improve practice. As a result, ReDWINE supports a system where patients get higher quality, safer, and more efficient care.

The ability to follow patients over time and across settings is particularly relevant for rehabilitation research since rehabilitation care often occurs across settings and over several months. This feature, however, has larger clinical and health system implications when it is useful to understand patient outcomes and health care utilization over time. For example, with payment models such as bundled payment and accountable care organizations, health systems receive a lump sum payment to manage patients over a specific period of time. Understanding the patient’s healthcare utilization trajectory over that time period can provide useful information to improve the quality and value of care.

The inclusion of NLP tools in ReDWINE enables rehabilitation researchers to study the effects of specific interventions to a much greater extent than was possible without NLP tools. Specific exercises and other interventions provided by rehabilitation providers are often only found within a free-text clinical note within a patient’s record. Therefore, simply examining CPT codes billed during a visit typically cannot provide the level of detail needed for a researcher to ascertain the effects of a specific intervention. For example, a physical therapist may bill for three 15-minute units of “therapeutic exercise” during a session. Using only structured data fields, it is typically not possible to know what exercises were performed. However, with NLP tools, researchers can extract data regarding the specific exercises and dosage (sets, repetitions, frequency, duration) provided, as shown in [7]. This will enable researchers to be much more precise in determining optimal treatment strategies for individuals seeking rehabilitation care.

The feature that ReDWINE leverages the OMOP CDM and the standardized vocabulary is critical for the rehabilitation EHR data because UPMC uses multiple EHR systems from different vendors across its various hospitals [9]. For example, different UPMC Home Health services use different EHR systems. Although UPMC is moving to an integrated Epic EHR system, integrating historical EHR data from multiple systems is still critical for rehabilitation research. ReDWINE is able to integrate and harmonize EHR data from multiple sources that follow different data formats and coding systems, and resolve any discrepancies between sources to ensure consistency and accuracy. It provides a flexible and scalable framework, enabling rehabilitation researchers to conduct observational studies, health services research, and quality improvement, as well as facilitating multi-site studies.

Although ReDWINE is primarily developed for observational studies and health services research, its potential of storing comprehensive EHR data could also be used to assess RCT feasibility [12] as well as to enhance recruitment of underrepresented populations in the local health system. As is well known, the low recruitment rate of underrepresented populations of racial and ethnic minorities in clinical trials remains a problem and has led to many health disparities. Using rich demographic information in ReDWINE from multiple sources has the potential to help identify underrepresented populations for RCTs.

4.1. Limitations

There are several limitations to ReDWINE. First, ReDWINE is still a prototype implemented using a small size of EHR dataset from a single data source. Hence, the system was not analyzed in the production context, with many users and a continuous stream of data. Second, ReDWINE currently relies on limited structured and unstructured data sources from EHRs, which may result in incomplete representation of the patient’s medical history. Many other EHR data, such as lab tests and image data, are note yet fully integrated into the system, potentially limiting the depth of insights that can be obtained. Third, while ReDWINE incorporates NLP techniques for extracting concepts from clinical texts, the accuracy and coverage of concept extraction may vary. Based a previous study, the OHNLP Toolkit achieved an accuracy of 71% [13] in a large multi-site study known as the national COVID cohort collaborative (N3C). Therefore, the outputs still contain potential inaccuracies or omissions in the extracted information. Fourth, ReDWINE’s performance heavily relies on the quality and consistency of the underlying data sources. Inconsistencies, missing data, or data entry errors can introduce biases or impact the reliability of the results obtained through the platform. Fifth, because we wanted to make the study accessible and reproducible, for all major data analysis and exploration tools, we relied on OHDSI framework and tools (e.g., ATLAS, HADES). Thus, the development and usage of custom, non-OHDSI tools was not thoroughly explored. Finally, since ReDWINE uses OMOP CDM, it may also have limitations that existed in the OMOP CDM.

4.2. Future Work

We have built a ReDWINE prototype using EHR data from a cohort of patients with stroke. The prototype has necessary functions to facilitate rehabilitation research. It is in its early stage to serve multiple users for research at the University of Pittsburgh SHRS. An immediate action item is to have several test users utilize ReDWINE for their observational studies and test the cohort definition and data analytics functions and seek their feedback to improve the prototype. To make ReDWINE a self-service tool for researchers, there are several items to be considered. First, we will work with the institution IRB to ensure that the investigator has proper IRB approvals prior to using ReDWINE. We will implement auditing and privacy settings to ensure the system is being used to its full potential by investigators. Second, we will develop a service and maintenance business model to ensure that the support team of ReDWINE is sustainable. Third, we will collaborate with the Neptune team to ensure the EHR data can be updated regularly. Fourth, we will develop several R wrappers for frequently used data retrieval functions for the HADES tool. This will make the data analytics tool more convenient for researchers without SQL experience. Fifth, in addition to the ongoing development of ReDWINE, we intend to undertake a rigorous usability study to assess its effectiveness and suitability for clinical researchers. This study will involve tracking various usability statistics, such as user satisfaction, ease of use, and task completion rates. Furthermore, we will closely monitor the frequency of use of ReDWINE by clinical researchers to gain insights into its practical applicability and acceptance within the research community. By conducting these assessments, we aim to gather scientific evidence regarding the usability and adoption of ReDWINE as a valuable tool for clinical research. Finally, with the rapid emergence of Artificial Intelligence (AI) models and techniques, it can be interesting to incorporate such capabilities into the data analytics tool in ReDWINE. While it is currently possible to automatically perform statistical analyses and training AI models (e.g., AutoML [14]), such functions have not been integrated into ReDWINE. Thus, implementing a pipeline for AI model training could be another avenue of exploration in the future work.

Authors' contributions

DO: implemented ReDWINE; analyzed the data; wrote the manuscript, BP: conceptualized the study; edited the manuscript, AS: edited the manuscript, AB: edited the manuscript, JF: edited the manuscript, SV: conceptualized the study; edited the manuscript, NC: retrieved the data; edited the manuscript, BM: retrieved the data, JCS: conceptualized the study; edited the manuscript, MJB: conceptualized the study; edited the manuscript, AD: conceptualized the study; edited the manuscript, ES: conceptualized the study; edited the manuscript, YW: designed and implemented ReDWINE, conceptualized the study; analyzed the data; wrote the manuscript. All authors read and approved the final manuscript.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors would like to acknowledge the OHDSI community for providing extensive documentation of the CDM and releasing the OMOP data analytics tools. This project was supported by the University of Pittsburgh School of Health and Rehabilitation Sciences and the National Institutes of Health through Grant Numbers UL1TR001857 and U24TR004111 (SV, NC, BM, JCS, MJB, YW). The Neptune data warehouse is supported by the National Center for Advancing Translational Sciences, The Center for Disease Control and National Institute for Occupational Safety and Health, NIH Common Fund and The Patient Centered Outcome Research Institute Clinical Research Network (UL1TR001857, RI-CRN-2020-006, OT2OD026554 and 5U24 OH009077 – to MJB, JCS, SV). The funders had no role in the design of the study, and collection, analysis, and interpretation of data and in preparation of the manuscript. The views presented in this report are not necessarily representative of the funder's views and belong solely to the authors.

Summary table

We developed a clinical research datamart, called ReDWINE (Rehabilitation Datamart With Informatics iNfrastructure for rEsearch), that transforms the rehabilitation-related electronic health record data collected from the a local health care system to the Observational Health Data Sciences and Informatics (OHDSI) Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) to facilitate rehabilitation research.

ReDWINE includes deployment of data visualization and data analytics tools to facilitate cohort definition and clinical data analysis.

ReDWINE provide text analytical capabilities at scale by implementing a high-throughput natural language processing pipeline.

The data quality assurance analysis indicates the electronic health record data were successfully transferred to the OMOP CDM in ReDWINE.

Three case studies were demonstrated the use of ReDWINE for rehabilitation research.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ijmedinf.2023.105144.

References

[1] [https://www.who.int/news-room/fact-sheets/detail/rehabilitation].

[2] T. Kroll, J. Morris, Challenges and opportunities in using mixed method designs in rehabilitation research, Arch. Phys. Med. Rehabilitat., 2009, 90:S11-S16.

[3] A. Callahan, N.H. Shah, J.H. Chen, Research and reporting considerations for observational studies using electronic health record data, Ann. Internal Med. 172 (2020) S79–S84.

[4] OHDSI: The Book of OHDSI: Observational Health Data Sciences and Informatics. OHDSI, 2019.

[5] S.N. Murphy, S. Visweswaran, M.J. Becich, T.R. Campion, B.M. Knosp, G.B. Melton-Meaux, L.A. Lenert, Research data warehouse best practices: catalyzing national data sharing through informatics innovation, vol. 29. pp. 581-584: Oxford University Press; 2022:581-584.

[6] M. Becich, Information management: moving from test results to clinical information, Clin. Leadership Manage. Rev: J. CLMA 14 (2000) 296–300.

[7] S.W. Shaffran, F. Gao, P.E. Denny, B.M. Aldhahwani, A. Bove, S. Visweswaran, Y. Wang, Extracting Physical Rehabilitation Exercise Information from Clinical Notes: a Comparison of Rule-Based and Machine Learning Natural Language Processing Techniques, arXiv preprint arXiv:230313466 2023.

[8] A. Wen, S. Fu, S. Moon, M. El Wazir, A. Rosenbaum, V.C. Kaggal, S. Liu, S. Sohn, H. Liu, J. Fan, Desiderata for delivering NLP to accelerate healthcare AI advancement and a Mayo Clinic NLP-as-a-service implementation, NPJ Digital Med. 2 (2019) 130.

[9] S. Visweswaran, B. McLay, N. Cappella, M. Morris, J.T. Milnes, S.E. Reis, J. C. Silverstein, M.J. Becich, An atomic approach to the design and implementation of a research data warehouse, J. Am. Medi. Informatics Associat. 29 (2022) 601–608.

[10] R.J. Cotton, R.L. Segal, B.A. Seamon, A. Sahu, M.M. McLeod, R.D. Davis, S. L. Ramey, M.A. French, R.T. Roemmich, K. Daley, Precision rehabilitation: optimizing function, adding value to health care, Arch. Phys. Med. Rehabilitat. 103 (2022) 1883–1884.

[11] M.A. French, R.T. Roemmich, K. Daley, M. Beier, S. Penttinen, P. Raghavan, P. Searson, S. Wegener, P. Celnik, Precision rehabilitation: optimizing function, adding value to health care, Arch. Phys. Med. Rehabilitat. 103 (2022) 1233–1239.

[12] N. Ahmadi, Y. Peng, M. Wolfien, M. Zoch, M. Sedlmayr, OMOP CDM Can Facilitate Data-Driven Studies for Cancer Prediction: A Systematic Review, Int. J. Mol. Sci. 23 (2022) 11834.

[13] S. Liu, A. Wen, L. Wang, H. He, S. Fu, R. Miller, A. Williams, D. Harris, R. Kavuluru, M. Liu, An open natural language processing development framework for EHR-based clinical research: a case demonstration using the national COVID cohort collaborative (N3C), arXiv preprint arXiv:211010780 2021.

[14] X. He, K. Zhao, X. Chu, AutoML: A survey of the state-of-the-art, Knowledge-Based Systems 212 (2021), 106622.