JAMIA Open, 5(1), 2022, 1–10
https://doi.org/10.1093/jamiaopen/ooab110
Research and Applications

Research and Applications

Generating real-world data from health records: design of a patient-centric study in multiple sclerosis using a commercial health records platform

Gillian Hanson[1], Tanuja Chitnis[2,3,4], Mitzi J. Williams[5], Ryan William Gan[6,*], Laura Julian[6], Kieran Mace[1], Jenny Chia[6], David Wormser[7,†], Michael Martinec[7], Troy Astorino[1], Noga Leviner[1], Pye Maung[1], Asif Jan[7,‡], and Katherine Belendiuk[6]

[1]PicnicHealth, San Francisco, California, USA, [2]Harvard Medical School, Harvard University, Boston, Massachusetts, USA, [3]Mass General Brigham Pediatric Multiple Sclerosis Center, Massachusetts General Hospital, Boston, Massachusetts, USA, [4]Translational Neuroimmunology Research Center and Brigham Multiple Sclerosis Center, Brigham and Women's Hospital, Boston, Massachusetts, USA, [5]Joi Life Wellness Group MS Neurology Center, Smyrna, Georgia, USA, [6]Genentech Inc., San Francisco, California, USA, and [7]F. Hoffmann-La Roche Ltd, Basel, Switzerland

*Present address: The Janssen Pharmaceutical Companies of Johnson & Johnson, San Francisco, California, USA.
†Present address: Novartis International AG, Basel, Switzerland.
‡Present address: Owkin, Inc., Basel, Switzerland.
Corresponding Author: Katherine Belendiuk, PhD, Genentech Inc., 1 DNA Way, South San Francisco, CA 94080, USA; belendiuk.katherine@gene.com

Received 22 March 2021; Revised 21 October 2021; Editorial Decision 3 December 2021; Accepted 13 December 2021

ABSTRACT

Objective: The FlywheelMS study will explore the use of a real-world health record data set generated by PicnicHealth, a patient-centric health records platform, to improve understanding of disease course and patterns of care for patients with multiple sclerosis (MS).

Materials and Methods: The FlywheelMS study aims to enroll 5000 adults with MS in the United States to create a large, deidentified, longitudinal data set for clinical research. PicnicHealth obtains health records, including paper charts, electronic health records, and radiology imaging files from any healthcare site. Using a large-scale health record processing pipeline, PicnicHealth abstracts standard and condition-specific data elements from structured (eg, laboratory test results) and unstructured (eg, narrative) text and maps these to standardized medical vocabularies. Researchers can use the resulting data set to answer empirical questions and study participants can access and share their harmonized health records using PicnicHealth's web application.

Results: As of November 24, 2020, more than 4176 participants from 49 of 50 US states have enrolled in the FlywheelMS study. A median of 200 pages of records have been collected from 14 different doctors over 8 years per participant. Abstraction precision, established through inter-abstractor agreement, is as high as 97.8% when identifying and mapping data elements to a standard ontology.

Conclusion: Using a commercial health records platform, the FlywheelMS study is generating a real-world, multimodal data set that could provide valuable insights about patients with MS. This approach to data collection and abstraction is disease-agnostic and could be used to address other clinical research questions in the future.

Key words: real-world data, health records, data abstraction, machine learning, multiple sclerosis

Lay Summary

Health records contain valuable information about patients and the care they receive in routine clinical practice; however, use of this data source in research is hindered by the difficulty of obtaining complete analyzable data sets. In the United States, health records for each patient are stored in paper and electronic formats across multiple healthcare providers. Furthermore, data must be extracted from health records before they can be analyzed, which is technically difficult for images and free text. In the first part of this paper, we describe how PicnicHealth, a commercial health records platform, collects health records on behalf of patients in any format and from all healthcare sites. Facilitated by tailored software tools and task-specific machine-learning models, data are extracted from health records by human experts in an efficient and precise manner. This enables patients to access and manage their health record data via a web application. The second part of this paper describes the design, rationale, and recruitment metrics for the ongoing FlywheelMS study, which is exploring whether an anonymized data set generated by PicnicHealth can improve our understanding of the disease course and patterns of care for patients with multiple sclerosis.

INTRODUCTION

Real-world data are data collected outside of a clinical trial setting.[1] These data, which can be obtained from a variety of sources, provide valuable insights about disease course and patient care in routine clinical practice.[1] One of the key strengths of real-world data is the ability to capture information on patients who are traditionally underrepresented in clinical trials, including those with comorbidities or complex treatment histories, the elderly, people from ethnic minority groups, and those living in rural communities.[2–4]

Among common sources of real-world data for clinical research (eg, health records, registries, and claims databases), health records are particularly valuable.[1] Health records can provide longitudinal data on large cohorts of patients in a time- and cost-effective manner.[5] Furthermore, the detailed clinical information captured in free-text notes and diagnostic reports written by clinicians,[6] such as the rationale behind treatment decisions, provides clinical context that is unavailable in claims databases or research registries. Health records also contain more complete data than claims databases and registries because these are susceptible to gaps caused by lapses in insurance coverage or changes in insurance provider,[7] and loss to follow-up,[8] respectively.

Despite the value of health records, their widespread use in research is hindered by the difficulty of obtaining complete analyzable data sets. In the United States, health records are collected and stored in both electronic and paper formats, and are siloed across multiple healthcare systems. Poor interoperability of electronic health record (EHR) systems further hampers efforts to share and standardize data.[9–12] Owing to these difficulties, researchers may rely on arrangements with EHR providers or hospital systems to obtain analyzable data. However, this approach limits data set completeness because data capture is restricted, for example, to care sites using a particular EHR system. Health record data sets may only include structured information (eg, values presented in tables) owing to the technical difficulty of extracting data from unstructured narrative text or medical images.[10] For these reasons, data collected outside of a specific EHR system and/or data from unstructured text may be missing from health record data sets.[10,11]

To overcome the issue of missing data in real-world data sets, the FlywheelMS study utilizes a health records platform developed by PicnicHealth. Health records are collected in any format from any US-based healthcare site and processed via the platform to produce a complete timeline of health records. Rather than aiming to automate perfect data abstraction for any possible record, PicnicHealth's novel approach uses task-specific machine learning and an extensive assortment of software tools to facilitate the processing of health records by expert human abstractors. This enables the research of large patient populations, including patients who are often underrepresented in clinical trials, and overcomes the common causes of real-world data missingness. Additionally, the efficient collection and abstraction of health record data via the platform allows patients to access and manage their health records using a web application, and facilitates the secure sharing of records with clinicians (Figure 1). The FlywheelMS study aims to establish whether a large-scale, deidentified, longitudinal data set created using the PicnicHealth platform can enable research in a broad population of patients with multiple sclerosis (MS).

MS is a chronic neurological condition of the brain and spinal cord.[13] The disease course and symptoms are highly variable, and use of, and response to, available disease-modifying treatments differ among individuals.[14,15] Therefore, the FlywheelMS study is an important opportunity to utilize longitudinal data from health records from patients with MS to better understand disease course, treatments, and the response of patients to different routine care strategies.

OBJECTIVE

This article provides an overview of the health record processing pipeline, which retrieves health records and abstracts data elements from structured and unstructured text to create a large, multimodal data set. The design and rationale of the FlywheelMS study, as well as the latest recruitment progress, are described and metrics for both the data abstraction process and the resulting data set are summarized. Finally, the potential implications of using this approach to gain real-world insights about patients and their disease are discussed.

MATERIALS AND METHODS

Health records platform

Generating an analyzable data set from health records involves several steps, including the retrieval of health records, abstraction of data, and quality control (Figure 2).

Patient authorization and record retrieval

First, individuals create a secure account, provide identifying information for record collection, consent to participate in the study, and electronically sign research and record release authorization forms.

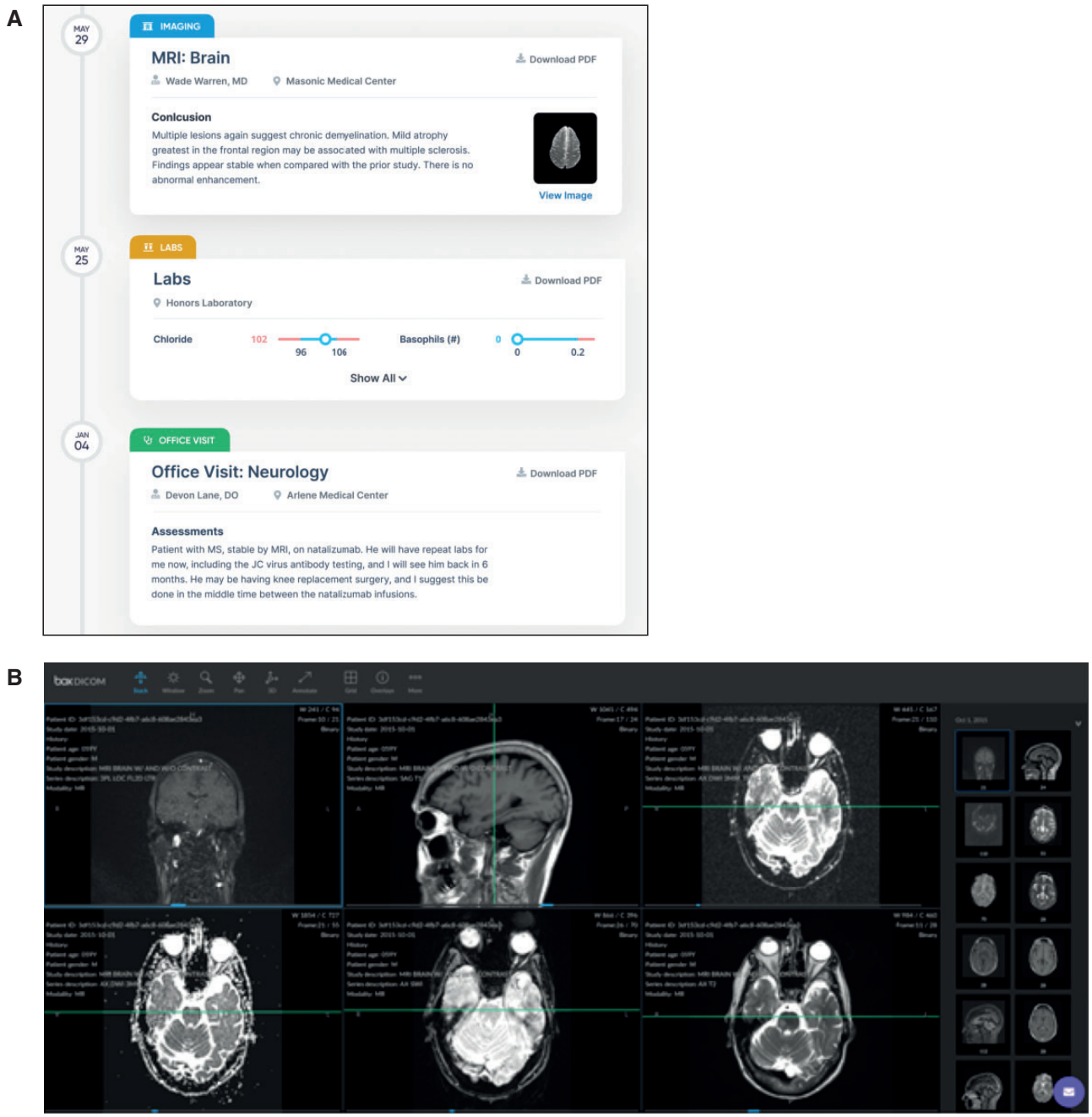

Figure 1. Web application interface showing (A) health record timeline for a patient with multiple sclerosis and (B) brain magnetic resonance imaging. An interactive demo is available at https://picnichealth.com/demo. *Abbreviations:* JC virus: John Cunningham virus; MRI: magnetic resonance imaging; MS: multiple sclerosis.

Patients are then prompted to provide details of their clinicians and the name and location of the healthcare facilities where they have received clinical care. These details are mapped to a proprietary database of more than 28 000 care sites, which is updated with new care sites as they are reported by patients. Contact instructions necessary to retrieve records from each facility are actively maintained within this database. Record retrieval requests are automatically generated and sent to the appropriate recipient based on experiences gleaned from prior requests to that facility. If the available information is insufficient for identification of a clinician or healthcare site, the patient is contacted to provide additional details. To date, health records have been obtained from 98.7% of healthcare facilities listed by users of PicnicHealth.

After sending the record retrieval request, receipt of the request is confirmed with the facility and proactively followed up until the records are received. Most requests are fulfilled within 4–6 weeks. The turnaround time is fastest for requests to facilities that have previously provided records for a patient, but can be significantly longer for a small portion of facilities depending on their staffing levels and internal processes. In 2019, 75% of records were received within 21 days. This fast turnaround time is a result of continual optimizations made to PicnicHealth's retrieval tools, processes, and training. For example, addition of an autocomplete feature to the facility search user interface improved access to, and maintenance of, contact instructions by more than 20%, as measured by database entry reuse, enabling critical information to be more consistently kept up to date.

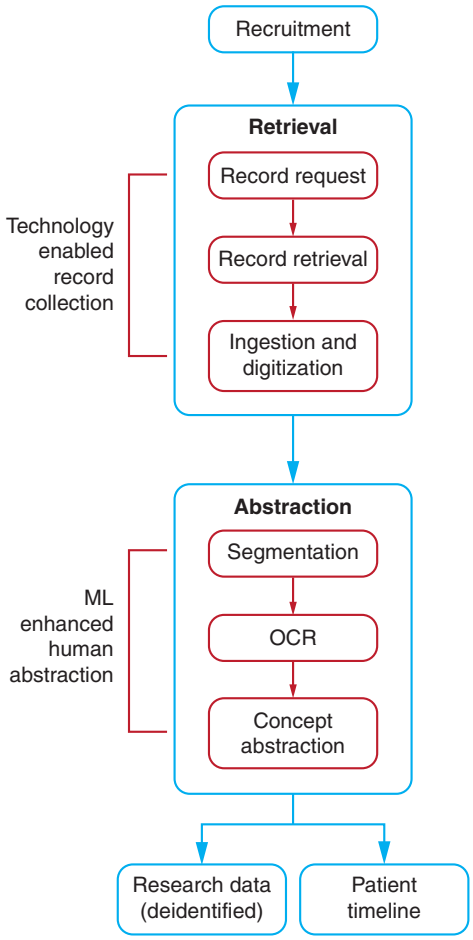

Figure 2. Data abstraction workflow for standard data elements. *Abbreviations:* ML: machine learning; OCR: optical character recognition.

To accommodate the disparate formats used to store health records, records are accepted from a diverse set of media transmitted via fax, postal mail, or email, including electronic health files, paper records, and digital media (eg, CDs, USB drives). All records are reviewed for quality and completeness (eg, missing pages, illegible text) and higher quality copies are requested if necessary. Health records are stored on encrypted servers and all physical copies of records are destroyed after record verification and digitization.

Data abstraction process for text-based health records

The data abstraction process transforms the contents of health records so they can be included in a structured analyzable data set. The resulting data set contains standard elements for all patients including demographic information, diagnoses, medications, vital signs, and laboratory results. They also contain condition-specific custom data elements. A specialized set of algorithms trained on PicnicHealth's existing health records corpus allows repetitive steps in the abstraction pipeline to be automated, making the process scalable to the order of tens of thousands of pages per day in 2020. On their own, state-of-the-art machine learning models are not yet sufficient to map data in arbitrary records to medical ontologies for research purposes.[16–18] Therefore, a novel human-in-the-loop approach has been adopted, which does not require machine learning to provide perfect abstraction for any possible record, but instead uses task-specific machine learning models to improve the efficiency of trained human abstractors. Data set accuracy does not suffer when machine learning models are imperfect because experts process data elements that are difficult for models and perform comprehensive quality checking. Together with sophisticated software systems, personnel management policies, and economies of scale, machine learning improvements increase the number of documents that the pipeline can process.

Document preprocessing

The first steps in the health record processing pipeline are to digitize paper records using optical character recognition (OCR), extract metadata associated with visits to clinicians (eg, date, type, care site, and clinician), and demarcate sections of records for downstream data abstraction. Section types include medication lists, assessments, treatment plans, problem lists, vital signs, laboratory results, and interpretations.

Each of these tasks is performed using bespoke software, which is designed so that human abstractors can quickly confirm, override, or supplement the outputs of data-specific (eg, laboratory results, conditions) machine learning models. Machine learning models, software and its user interfaces, and management tools are continually refined. The impact of any pipeline improvement is measured both in terms of abstraction time and inter-abstractor agreement on the same task, which is targeted to remain above 98% agreement for any improvement to be adopted.

An example of a machine learning model that has been refined is the OCR model, which is based on the U-Net architecture for image segmentation[19] with sub-networks in the style of DenseNet[20] to process pixel-level information and connectionist temporal classification layers[21] to make word-level predictions.

Abstraction of standard data elements

Similar to other steps in the pipeline, abstraction tasks for each type of standard data element are implemented using bespoke software, assisted by the outputs of task-specific machine learning models. OCR-corrected words within each section of the health record provide the input for the data abstraction process. Models output an assignment of words to a structured concept, as well as an ontology code and additional values, such as numeric values for lab measurements. These outputs are automatically populated in the software application's user interface, and then confirmed, corrected, and supplemented by human abstractors working in teams specialized by concept type (eg, medication, laboratory test). Abstraction difficulty varies by concept type and is driven by lexical variations, the need for domain knowledge, and the presence of implicit information, such as whether a medication was administered to a patient according to a document's heading. At present, vital sign abstraction is the most efficient, with laboratory result abstraction requiring 2.5 times more manual processing time and medications requiring 7 times more manual processing time than vital signs.

Within each section of a report, words are mapped to structured concepts using standard medical ontology systems, including RxNorm, International Classification of Diseases, Ninth and Tenth Revision, Clinical Modification (ICD-9-CM, ICD-10-CM), Logical Observation Identifiers Names and Codes (LOINC), Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT), and Observational Medical Outcomes Partnership Common Data Model (OMOP-CDM). Each concept is labeled with information pertinent to its data type. For instance, the LOINC concept code for a laboratory test result is labeled with the value, unit, and reference range. In

cases when a concept is not mappable to a standard ontology (eg, emerging diseases such as SARS-CoV-2 or experimental therapies), internally developed custom ontologies are used.

The integrity of all source documentation is maintained during this process. If the source pixels contain a misspelling, neither the OCR model nor any subsequent human review corrects these errors. The abstraction models, however, are generally robust to spelling errors and morphological variants. Edge cases not caught by models are still captured during the structuring process because each model is positioned as an accelerator to human abstractors who ensure correctness.

Abstraction of custom data elements

Custom data elements and concept synonyms are developed to extract information from structured and narrative text relating to a specific research question or disease area. These are defined according to publicly available sources, such as the Unified Medical Language System (UMLS) Metathesaurus, SNOMED CT relationships, as well as expert clinical consultation. Custom abstraction follows a workflow that includes additional training for abstractors and a software interface designed for these low volume, high complexity data elements. One such example is the identification of possible MS disease progression based on records of assistive device use, such as a wheelchair, although this is not recorded as frequently as the use of medications. Once identified, custom element mentions are labeled with selected additional fields to provide necessary context (eg, subject, temporality, negation).

Quality control

Multiple processes are used to ensure high accuracy during the abstraction process. All abstractors complete an initial training program for the tasks assigned to their role, followed by an evaluation and practical teaching period. During this latter stage, inter-abstractor agreement on task outputs is used to identify knowledge gaps and trigger example-based coaching. Training is also provided for specific disease areas to improve decision-making when abstracting data elements from ambiguous text.

All entity-level data are verified in a quality control stage, during which prior work is confirmed by a second human abstractor. Complex data elements or scenarios that are not described in standard protocols are escalated to a senior team lead, with doctors and specialty nurses available for further escalations. Inter-abstractor agreement studies are regularly performed as an additional quality assurance metric.

Ensuring visit completeness

In addition to the measures described above, which are designed to increase confidence related to data abstracted from each encounter, explanation of benefits data are used to confirm that records from all available encounters are being collected. A custom workflow is used to retrieve insurance information provided during an optional step in the participant enrollment process. The dates of visits known to insurers are abstracted from the explanation of benefits data and compared against visit dates for records that have been retrieved and abstracted. This information is used both to monitor retrieval effectiveness and to generate additional record requests, once confirmed by direct patient outreach via email or telephone.

Data storage and security

Data abstracted from text-based health records are stored in an OMOP-CDM-derived format[22] and medical images in the Digital Imaging and Communications in Medicine (DICOM) file format on encrypted servers. Personally identifiable information is removed from research data sets. The system is Health Insurance Portability and Accountability Act (HIPAA) security rule compliant[23] and Health Information Trust Alliance (HITRUST) certified.

FlywheelMS study

For a chronic and heterogeneous disease such as MS, a longitudinal multimodal health record data set may provide unique value to improve our understanding of disease course and treatments. As such, the FlywheelMS study is aiming to collect and abstract health record data from 5000 participants with MS in the United States using the PicnicHealth platform.

Ethics

The FlywheelMS study protocol has been approved by IntegReview (Austin, TX, USA), a central institutional review board. Deidentified research data sets are securely transferred from PicnicHealth's servers to the FlywheelMS researchers using end-to-end encryption.

Participants

To be eligible, individuals must have a self-reported diagnosis of MS confirmed by the presence of an MS diagnosis in their health records, provide informed consent for health record retrieval and abstraction, and be at least 18 years old at the time of consent. Individuals are excluded if they fail to list their clinicians or have no US-based health records.

Participants provide their informed consent electronically via the web application and have the right to withdraw their consent at any time and for any reason. If a participant withdraws their consent, data that has already been retrieved and structured at that time is included in the research data set; however, no further records are retrieved for the study after the time that the participant withdraws.

Recruitment

FlywheelMS aims to recruit 5000 participants in the United States through established partnerships with advocacy groups and providers, including organizations supporting people of color and those from disadvantaged socioeconomic backgrounds, and via a range of patient-facing channels, including social media platforms, conferences, email lists, websites, and physical media.

Individuals with MS are directed to the FlywheelMS study website (https://flywheel.ms/), which provides information on the study and the health records platform. FlywheelMS researchers are available to answer additional questions by telephone or the email address provided on the study landing page. To join the FlywheelMS study, individuals complete a series of steps on the study enrollment webpage (Figure 3). Participants can access their health records via the web application without charge for the duration of the study and are able to download copies of their records.

Data collection

Data are collected retrospectively from existing health records, as well as prospectively when new health records are generated during clinical care. It is anticipated that data would typically be available for a minimum of 7 years prior to participant enrollment, owing to US data retention regulations. For the current study, data are collected prospectively for up to 5 years after enrollment. PicnicHealth identifies new healthcare visits by regularly contacting patients' current providers, by encouraging patients to input new clinician or

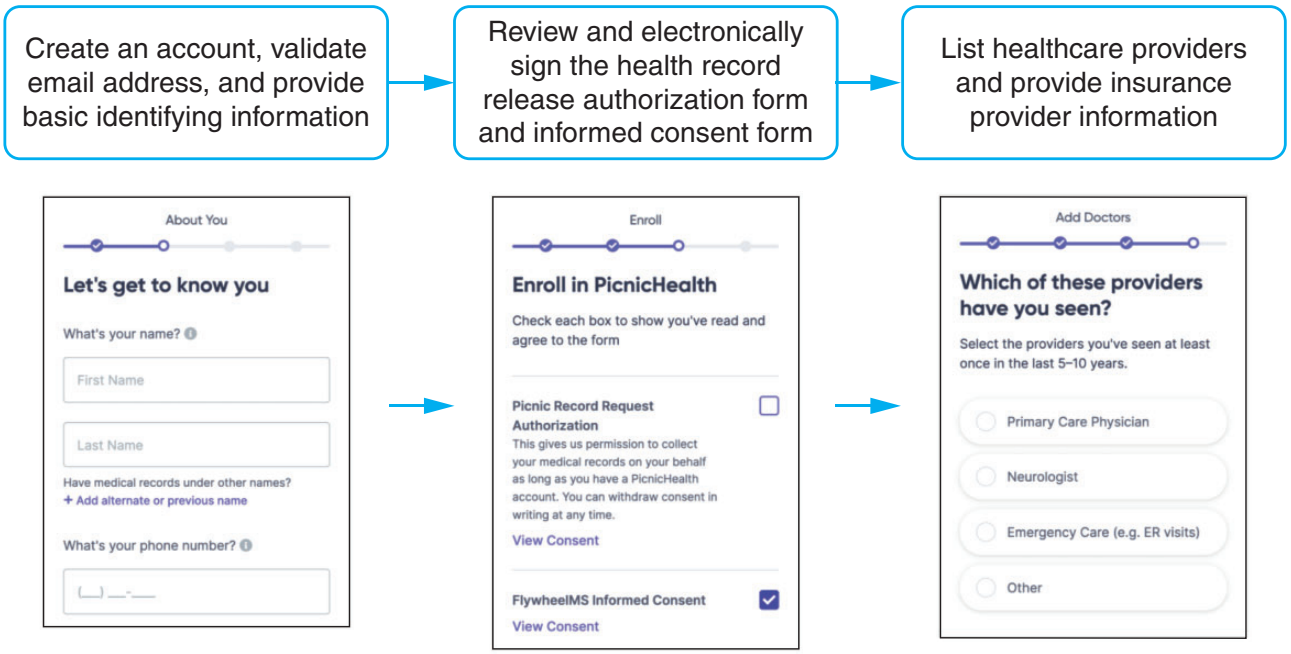

Figure 3. Enrollment process for FlywheelMS. All patients must complete the following steps to join the study. The informed consent form can be viewed at https://flywheel.ms/informed-consent.

visit information on the platform, and by identifying unknown visits from insurance data. Information relating to a new care site visit is added to the platform 2.6 months after the visit date on average. Patients who are missing reports related to neurology visits, magnetic resonance imaging (MRI) scans of the brain and/or spine, and internal medicine or family medicine visit reports are funneled into an additional email outreach campaign to solicit information about possible missing clinicians or records because it would be improbable that a patient with MS would not have engaged with these services during the routine course of care in the United States.

PicnicHealth's records processing pipeline abstracts both standard and custom data elements for the FlywheelMS study. Custom data elements included in FlywheelMS are designed to extract MS-specific variables, including MS subtype, treatment details, and indicators of MS relapse and progression (Table 1; Supplementary Material S1). FlywheelMS also captures brain MRI DICOM files for potential quantification.

RESULTS

Data abstraction performance metrics

OCR performance

Using a randomly held-out test set of 1000 records, PicnicHealth's OCR model averaged 8.3 mistakes per 100 words, a substantial improvement on the Google Cloud Vision API, which averaged 17.5 mistakes per 100 words on the same records. After manual review of OCR output, the error rate was 4 mistakes per 10 000 words based on disagreement with a second reviewer.

Precision of custom data element abstraction

The end-to-end precision of data abstraction was quantified based on agreement between two medically trained abstractors. Agreement, including both identification-correctness and ontology coding, was 99.5% for abstraction of MS diagnosis and subtype, 95.8% for abstraction of MS disease-modifying therapies, and 97.8% for potential indicators of MS relapse.

FlywheelMS study metrics

Recruitment progress

As of November 24, 2020, FlywheelMS has recruited 4176 participants with MS across the United States, including participants from 49 of 50 states (Figure 4).

Data set metrics

The median duration of data abstracted per participant is 8 years, up to a maximum of 48 years. The data set includes a median of 200 pages of health records, 14 different doctors, 4 hospital or emergency department visits, and 17 outpatient visits per participant (Table 2).

The median age of participants enrolled in FlywheelMS is 52 years. Participants are predominantly female (80.0%) and white (61.3%) (Table 3). Abstraction of race from EHRs is consistent with previous literature with regard to missingness.[24,25]

DISCUSSION

FlywheelMS is an ongoing study that has started to yield a large-scale, analyzable data set of health records from patients with MS. This data set will be used to address important questions relating to longitudinal disease course and the response of patients to different care strategies.

Health records are an underutilized resource in clinical research, owing to difficulties in accessing records, harmonizing data, and abstracting the depth of data contained within unstructured narrative text. Prior approaches to obtaining health records from individual clinicians or EHR systems fall short, missing visits occurring at other healthcare sites. Furthermore, technical barriers make it difficult for small-scale efforts to capture data that are not already structured and mapped to coded ontologies.

In contrast, PicnicHealth's patient-centric approach to health record retrieval allows records to be obtained from any healthcare site or type of facility, and in any format. All clinical notes and imaging

Table 1. Data elements abstracted from health records in the FlywheelMS study

Data element	Details	
Standard data elements		
Demographics	Sex	
	Month and year of birth	
	Race/ethnicity	
Visit information	Visit date	
	Visit type (eg, outpatient)	
	Specialty (eg, neurology)	
	Care site	
	Performing clinician	
Medications	Drug name/ingredient	
	Strength	
	Dose form	
	Start/stop date	
Medical conditions	Visit diagnoses	
	Problem lists	
	Start/stop dates	
Laboratory tests	Test name	
	Value	
	Units	
	Reference range	
Procedures	Procedure name	
	Care site	
	Performing clinician	
Imaging	Modality	
	Body part	
	Care site	
	Performing clinician	
Immunizations	Vaccine name	
	Care site	
Vitals	Vital sign test (eg, blood pressure)	
	Value	
	Unit	
Custom data elements		
MS diagnosis	"Multiple sclerosis"	Start date, negation[b]
	MS subtypes[a]	Start date, negation[b]
MS treatment	Disease-modifying therapies[a]	Start date, end date, temporality,[c] stop reason, route of administration
MS relapse	"MS relapse"	Start date, end date, negation,[b] temporality[c]
	"Optic neuritis"	Start date, end date, negation,[b] temporality[c]
	Corticosteroids[a]	Start date, end date, temporality,[c] route of administration
MS progression	"MS progression"	Start date, negation[b]
	Quantitative measures of progression (eg, EDSS, MACFIMS)[a]	Date performed, with assistance
	Assistive devices[a]	Start date, end date, negation,[b] temporality[c]
	"On permanent disability"	Start date, end date, negation[b]
Brain MRI quantification (planned)	Number and volume of existing lesions	
	Number and volume of new lesions	
	Changes in volume of lesions	
	Changes in whole brain, gray matter, and white matter volume	

Abbreviations: EDSS: Expanded Disability Status Scale; MACFIMS: Minimal Assessment of Cognitive Function in MS; MRI: magnetic resonance imaging; MS: multiple sclerosis.

[a]Further details are provided in Supplementary Material S1.

[b]Positive, possible, or negative.

[c]Past or current.

files are retrieved, and data elements are abstracted from both structured and narrative text using codes that can be customized to study objectives. Furthermore, the platform allows for both retrospective and prospective record collection. This approach produces a large-scale, longitudinal data set that has the potential to provide more detail and longer follow-up than site- or system-based data sets.

Existing automated natural language processing and machine learning approaches for health record data abstraction are not sufficiently accurate for clinical research. For extraction of medication information from free-text notes, current state-of-the-art systems fail to reach the minimum benchmark for clinical utility (F_1-score > 0.90–0.95), with F_1-scores ranging from 0.752 to 0.864.[16–18] Such

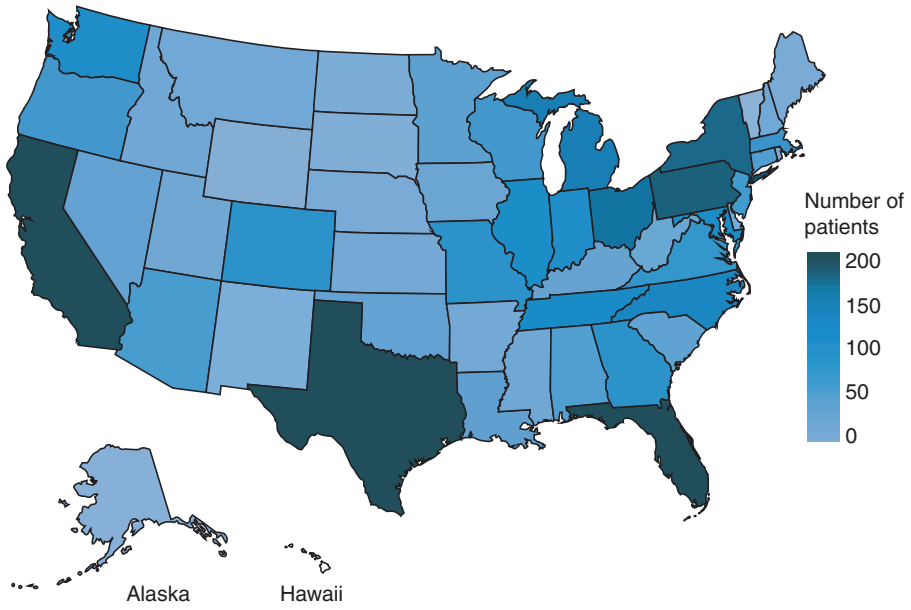

Figure 4. Choropleth map of patient recruitment.

Table 2. Data abstraction metrics for FlywheelMS

Metrics per participant	Median (Q1, Q3)
Pages	200 (90, 417)
Doctors	14 (7, 27)
Outpatient visits	17 (8, 37)
Neurology outpatient visits	9 (4, 16)
Hospital or ED visits	4 (2, 10)
MRI DICOM studies[a]	5 (3, 8)
MRI DICOM series[a]	52 (28, 92)
MRI DICOM slices (SOP)[a]	1823 (930, 3385)
Years of data	8 (4, 13)
Measurement entities	319 (134, 660)
Condition entities	152 (67, 331)
Drug entities	212 (87, 478)

Abbreviations: DICOM: Digital Imaging and Communications in Medicine; ED: emergency department; MRI: magnetic resonance imaging; SOP: service object pair; Q1: 25th percentile; Q3: 75th percentile.

[a]Among patients with available MRI data.

Table 3. Patient population metrics for FlywheelMS

Population metrics	
Age, years	
Median (Q1, Q3)	52 (42, 59)
Sex, %	
Male	19.5
Female	80.0
Missing	0.5
Race, %	
White	61.6
Black	6.3
Other	0.7
Missing	31.4

Abbreviations: Q1: 25th percentile; Q3: 75th percentile.

scores yield a significant number of errors that undermine any conclusions drawn from the structured data. In contrast, our approach overcomes this challenge by utilizing machine learning to solve a more tractable problem: improving the efficiency of an existing pipeline while retaining high levels of precision. This approach also brings to bear a richer set of tools to scale the processing of health records. Currently, PicnicHealth abstracts data from tens of thousands of pages per day, indicating the success of human-in-the-loop machine learning for generating real-world data at scale.

Researchers can leverage structured, longitudinal data sets extracted from health records for clinical research in a variety of disease areas. As a proof-of-principle use case, the FlywheelMS study will create a large-scale data set of health records from patients with MS; however, in the future, this approach could be expanded to study the disease course and routine care of patients with other diseases. At present, FlywheelMS includes data covering 8 years on average per participant. Analysis of these longitudinal data could provide novel insights into the clinical profile of patients with MS before disease onset, MS disease course, and the comparative effectiveness of available treatments. It is anticipated that the FlywheelMS data set will evolve over time with addition of new site visits and be enriched with other real-world data sources, expanding the potential research applications of the data. For example, the addition of structured radiological data extracted from brain MRI images could be combined with clinical data for the development of diagnostic, prognostic, and predictive models.[26,27]

FlywheelMS is a real-world health record data abstraction study and, as such, has limitations that are inherent to the source data and study design. Data stored in health records, particularly data recorded several years ago, can be incomplete and non-standardized, while some health records will have been lost, destroyed, or are irretrievable. Owing to differences in documentation processes between clinicians, an evolving understanding or interpretation of a patient's medical history or diagnosis over time, and clinician error, data stored in different records could have contradictory implications,

which are challenging to interpret. Recruitment is biased toward English speakers, individuals treated in major health centers, those who are engaged with the wider MS community, and those who are computer literate; therefore, the study cohort may not be representative of all patients with MS, despite efforts to recruit patients from diverse backgrounds. There are also limitations that arise from the data abstraction process, including the potential for algorithmic and human error. Finite synonym dictionaries for concepts may be unable to accommodate all cases and linguistically complex context may not be captured adequately by a coded data set. Despite these limitations, this patient-centric approach overcomes several key shortcomings of previous health record studies and provides the foundation for deriving clinically relevant insights at a population level about disease and the care of patients in the real world.

CONCLUSION

Using a commercial health records platform, the FlywheelMS study will produce a large, longitudinal, multimodal data set with the aim of providing novel insights and answering important clinical questions relating to disease course and patterns of care for patients with MS in the United States. This patient-centric approach has the potential to overcome the challenges that limit the use of health records for research and could be applicable to other disease areas.

FUNDING

This study was funded by F. Hoffmann-La Roche Ltd.

AUTHOR CONTRIBUTIONS

AJ was involved in design of the data processing pipeline on behalf of F. Hoffmann-La Roche, and provided feedback to PicnicHealth on data model implementation. NL was involved in study design and protocol development. TA was involved in study design, protocol development, and data model development. GH was involved in study design, data model development, and manuscript development. PM was involved in study design and data model development. KM was involved in data analysis and manuscript development. KB was responsible for study conceptualization, protocol development, study execution, development of the data model, and manuscript development. RG provided summary statistics. All authors critically reviewed the manuscript and approved the final version for submission.

SUPPLEMENTARY MATERIAL

Supplementary material is available at *JAMIA Open* online.

ACKNOWLEDGEMENTS

Under the direction of the authors and funded by F. Hoffmann-La Roche Ltd, Dr C. Evans, of Oxford PharmaGenesis, Cardiff, UK, provided writing assistance for this publication. The authors thank Zongqi Xia, Kottil Rammohan, and Stephen J. Tarsa for their review of the manuscript.

CONFLICTS OF INTEREST STATEMENT

NL and TA are co-founders of PicnicHealth. GH, KM, and PM are employees of PicnicHealth. NL, TA, GH, KM, and PM hold stock options in PicnicHealth. JC, KB, LG, and RG are employees of Genentech, Inc. KB is a shareholder of F. Hoffmann-La Roche Ltd and Takeda Pharmaceutical Company Ltd. MM is an employee of F. Hoffmann-La Roche Ltd. AJ was an employee of F. Hoffmann-La Roche Ltd while involved in the study and is a current employee of Owkin, Inc. DW was an employee of F. Hoffmann-La Roche Ltd while involved in the study and is a current employee of Novartis International AG. MJW has received consulting and/or speaking fees from Alexion, Biogen Idec, Sanofi Genzyme, Genentech, Novartis Pharmaceuticals, AbbVie, Bristol Myers Squibb, and EMD Serono.

DATA AVAILABILITY

The data underlying this article are subject to an embargo of 6 years from the publication date of the article to allow for completion of prospective data collection and data analysis for the FlywheelMS study. Once the embargo expires, the data will be available through an application process managed by PicnicHealth and limited to researchers working with recognized academic or research organizations.

REFERENCES

1. Makady A, de Boer A, Hillege H, *et al.*; on behalf of GetReal Work Package 1. What is real-world data? A review of definitions based on literature and stakeholder interviews. *Value Health* 2017; 20 (7): 858–65.
2. Kim SH, Tanner A, Friedman DB, *et al.* Barriers to clinical trial participation: a comparison of rural and urban communities in South Carolina. *J Community Health* 2014; 39 (3): 562–71.
3. Nallamothu BK, Hayward RA, Bates ER. Beyond the randomized clinical trial. *Circulation* 2008; 118 (12): 1294–303.
4. Kwiatkowski K, Coe K, Bailar JC, *et al.* Inclusion of minorities and women in cancer clinical trials, a decade later: have we improved? *Cancer* 2013; 119 (16): 2956–63.
5. Casey JA, Schwartz BS, Stewart WF, *et al.* Using electronic health records for population health research: a review of methods and applications. *Annu Rev Public Health* 2016; 37: 61–81.
6. Hayrinen K, Saranto K, Nykanen P. Definition, structure, content, use and impacts of electronic health records: a review of the research literature. *Int J Med Inform* 2008; 77 (5): 291–304.
7. DeVoe JE, Gold R, McIntire P, *et al.* Electronic health records vs Medicaid claims: completeness of diabetes preventive care data in community health centers. *Ann Fam Med* 2011; 9 (4): 351–8.
8. Solberg TK, Sørlie A, Sjaavik K, *et al.* Would loss to follow-up bias the outcome evaluation of patients operated for degenerative disorders of the lumbar spine? *Acta Orthop* 2011; 82 (1): 56–63.
9. Oye KA, Jain G, Amador M, *et al.* The next frontier: fostering innovation by improving health data access and utilization. *Clin Pharmacol Ther* 2015; 98 (5): 514–21.
10. Berger ML, Curtis MD, Smith G, *et al.* Opportunities and challenges in leveraging electronic health record data in oncology. *Future Oncol* 2016; 12 (10): 1261–74.
11. Cowie MR, Blomster JI, Curtis LH, *et al.* Electronic health records to facilitate clinical research. *Clin Res Cardiol* 2017; 106 (1): 1–9.
12. Daniel G, Silcox C, Bryan J, *et al.* Characterizing RWD quality and relevancy for regulatory purposes; 2018. https://healthpolicy.duke.edu/sites/default/files/atoms/files/characterizing_rwd.pdf. Accessed June 23, 2020.
13. Gross HJ, Watson C. Characteristics, burden of illness, and physical functioning of patients with relapsing-remitting and secondary progressive multiple sclerosis: a cross-sectional US survey. *NDT* 2017; 13: 1349–57.
14. Ziemssen T, Kern R, Thomas K. Multiple sclerosis: clinical profiling and data collection as prerequisite for personalized medicine approach. *BMC Neurol* 2016; 16: 124.

15. National Multiple Sclerosis Society. Disease-modifying therapies for MS; 2019. http://www.nationalmssociety.org/nationalmssociety/media/msnationalfiles/brochures/brochure-the-ms-disease-modifying-medications.pdf. Accessed June 23, 2020.
16. Patrick J, Li M. High accuracy information extraction of medication information from clinical notes: 2009 i2b2 medication extraction challenge. *J Am Med Inform Assoc* 2010; 17 (5): 524–7.
17. Guzman B, Metzger I, Aphinyanaphongs Y, *et al.* Assessment of Amazon Comprehend Medical: medication information extraction. *arXiv.org* 2020; arXiv:2002.00481 (preprint: not peer reviewed).
18. Tao C, Filannino M, Uzuner Ö. Prescription extraction using CRFs and word embeddings. *J Biomed Inform* 2017; 72: 60–6.
19. Ronneberger O, Fischer P, Brox TU. net: convolutional networks for biomedical image segmentation. In: Proceedings of the 18th International Conference on Medical Image Computing and Computer-Assisted Intervention; Munich, Germany; 2015: 234–41.
20. Huang G, Liu Z, Van Der Maaten L, *et al.* Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; Honolulu, HI; 2017: 4700–8.
21. Graves A, Fernández S, Gomez F, *et al.* Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In: Proceedings of the 23rd International Conference on Machine learning; Pittsburgh, PA; 2006: 369–76.
22. Blacketer C. Common data model; 2020. https://ohdsi.github.io/TheBookOfOhdsi/CommonDataModel.html. Accessed June 23, 2020.
23. Health insurance portability and accountability act of 1996. *Public Law* 1996; 104: 191.
24. Lee SJ, Grobe JE, Tiro JA. Assessing race and ethnicity data quality across cancer registries and EMRs in two hospitals. *J Am Med Inform Assoc* 2016; 23 (3): 627–34.
25. Lee W-C, Veeranki SP, Serag H, *et al.* Improving the collection of race, ethnicity, and language data to reduce healthcare disparities: a case study from an academic medical center. *Perspect Health Inf Manag* 2016; 13: 1–11.
26. Larue RT, Defraene G, Ruysscher DD, *et al.* Quantitative radiomics studies for tissue characterization: a review of technology and methodological procedures. *Br J Radiol* 2017; 90 (1070): 20160665.
27. Gillies RJ, Kinahan PE, Hricak H. Radiomics: images are more than pictures, they are data. *Radiology* 2016; 278 (2): 563–77.