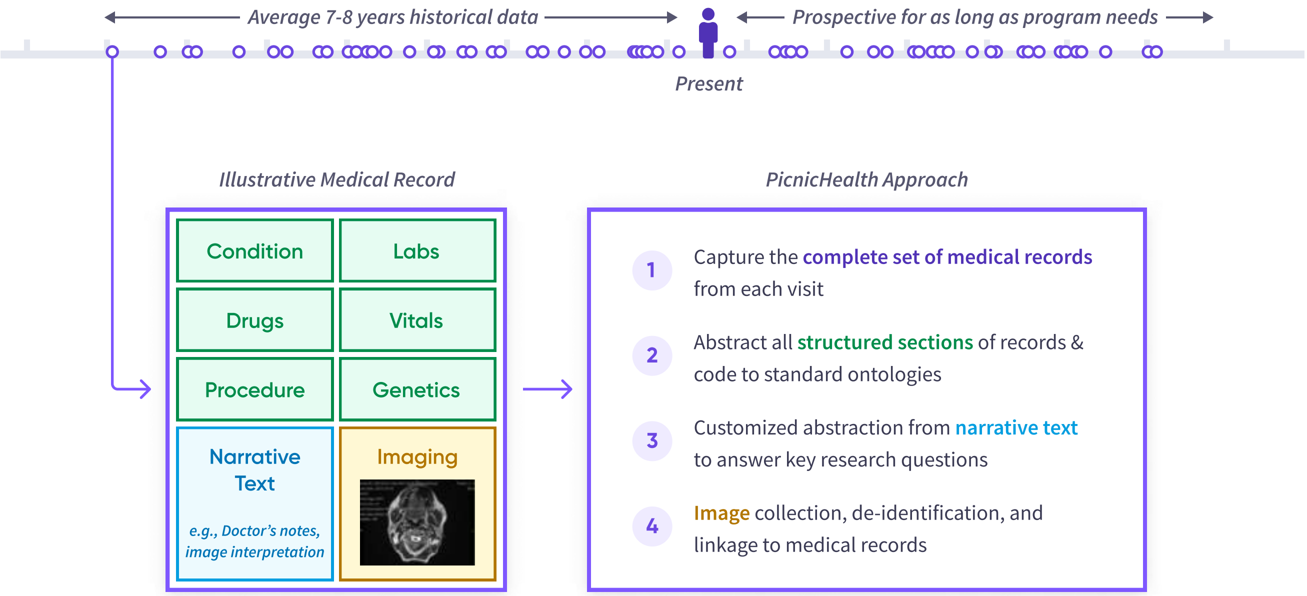
PicnicHealth

How PicnicHealth Generates Real-World Data from Medical Records

Overview

PicnicHealth partners with life science companies, academic research institutions, and health systems organizations in applying real world evidence for the advancement of patient care and treatment. We produce high quality multimodal datasets that cover a patient's entire medical journey and are more comprehensive and customizable than existing sources of real-world data.

Our data includes clinical information derived from structured and narrative texts, DICOM images, medical and pharmacy claims, and patient reported outcomes (PROs). On average, PicnicHealth has seven to eight years of retrospective data for each patient cohort. To date, we have collected over 3 million pages of medical records from over 90,000 different providers in the U.S.


Average 7-8 years historical data
Prospective for as long as program needs
Present
Illustrative Medical Record
Condition
Labs
Drugs
Vitals
Procedure
Genetics
Narrative Text
e.g., Doctor's notes, image interpretation
Imaging
PicnicHealth Approach

1. Capture the **complete set of medical records** from each visit
2. Abstract all **structured sections** of records & code to standard ontologies
3. Customized abstraction from **narrative text** to answer key research questions
4. **Image** collection, de-identification, and linkage to medical records

Figure 1 - PicnicHealth provides structured longitudinal data from all sections of the medical record, along with DICOM images.

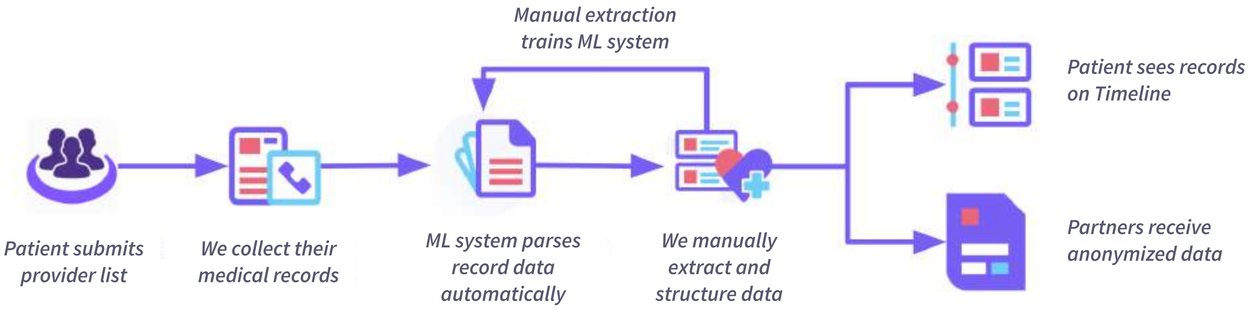
PicnicHealth has built a data processing pipeline that can collect records and extract data at scale. The process starts with recruiting and consenting patients, then reaching out to providers anywhere in the U.S. on behalf of patients. After records are retrieved, a PicnicHealth proprietary machine-learning (ML) system reads and parses clinical data automatically, aligns clinical concepts to standard medical ontologies, and experienced human annotators review all machine predictions to ensure data integrity. Patients benefit by gaining access to all their digitized medical records in one place, and research partners are able to use PicnicHealth's customized and de-identified data for real world evidence generation.

Human Subjects Review and Informed Consent

All PicnicHealth patient cohorts are developed under the guidance of an independent Institutional Review Board (IRB) which approves the study protocol and all patient-facing material including informed consent forms, patient-reported outcome instruments, etc.

De-identification of Personally Identifiable Information (PHI)

All personally identifiable information is de-identified during data processing and export. PicnicHealth follows Health Insurance Portability and Accountability Act (HIPAA) guidance in determining which data elements are identifiable, only de-identified data are exported to research partners [Appendix]. Individual patient records are only linked by system generated user IDs that do not contain any PHI. Certain data elements such as a patient's name, address, and photographs are never transmitted. Identifiable information are also removed from DICOM images (see DICOM Images section for details).



Manual extraction trains ML system
Patient submits provider list
We collect their medical records
ML system parses record data automatically
We manually extract and structure data
Patient sees records on Timeline
Partners receive anonymized data

Figure 2 - The data extraction process at PicnicHealth leverages both machine learning predictions and trained human reviewers.

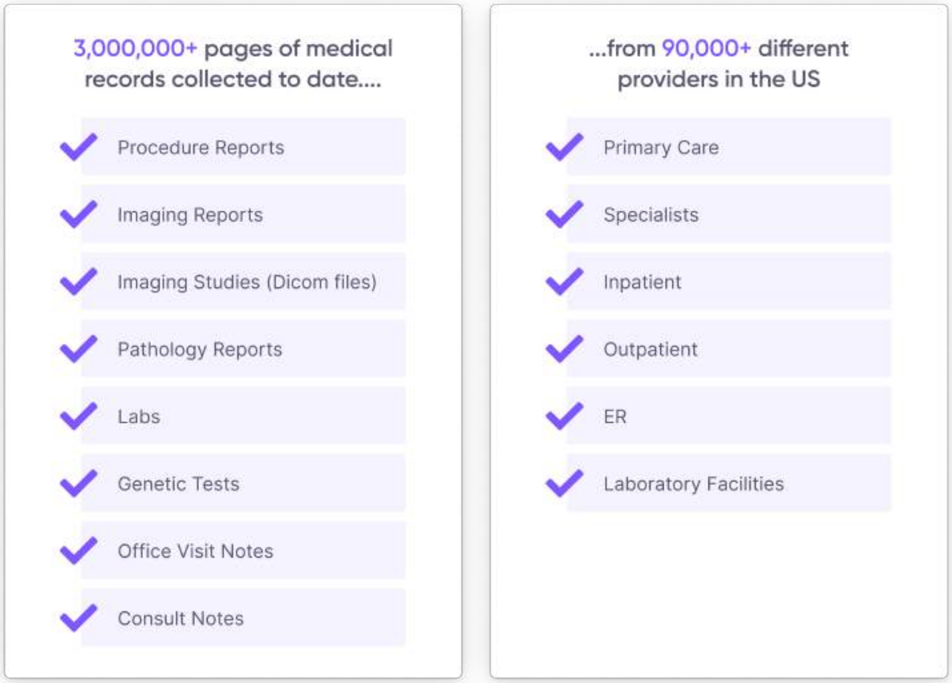
Recruitment and Record Retrieval

PicnicHealth's record retrieval process is targeted and comprehensive. We work with research partners to define patient cohorts based on the disease areas of interest and customize patient recruitment based on each study's qualification criteria. Patients are recruited from across the U.S. through advocacy groups, provider site-based partnerships, and social media campaigns. Once a patient signs up for the PicnicHealth platform and provides informed consent, PicnicHealth reaches out to those providers or facilities listed by the patient to gather all medical records on the patient's behalf.

Data collected includes any clinical documents, diagnostic and lab reports, and DICOM images. We enhance completeness of a patient's engagement with the healthcare system by:

- Mining records received for mentions of other providers or facilities
- Cross referencing administrative claims data to query for missing visits
- Modeling anticipated healthcare utilization for a given condition and patient and reaching out to the patient if unanticipated gaps in care exist
- Continuing prospective record collection throughout program enrollment

Medical records are received in any format used by the transmitting facility or provider, meaning that we accept paper records, faxes, secure web upload of PDF files, FHIR API, etc. PDF and paper documents are processed through our optical character recognition (OCR) system which generates searchable text from the original images.



3,000,000+ pages of medical records collected to date....

✓ Procedure Reports
✓ Imaging Reports
✓ Imaging Studies (Dicom files)
✓ Pathology Reports
✓ Labs
✓ Genetic Tests
✓ Office Visit Notes
✓ Consult Notes

...from 90,000+ different providers in the US

✓ Primary Care
✓ Specialists
✓ Inpatient
✓ Outpatient
✓ ER
✓ Laboratory Facilities

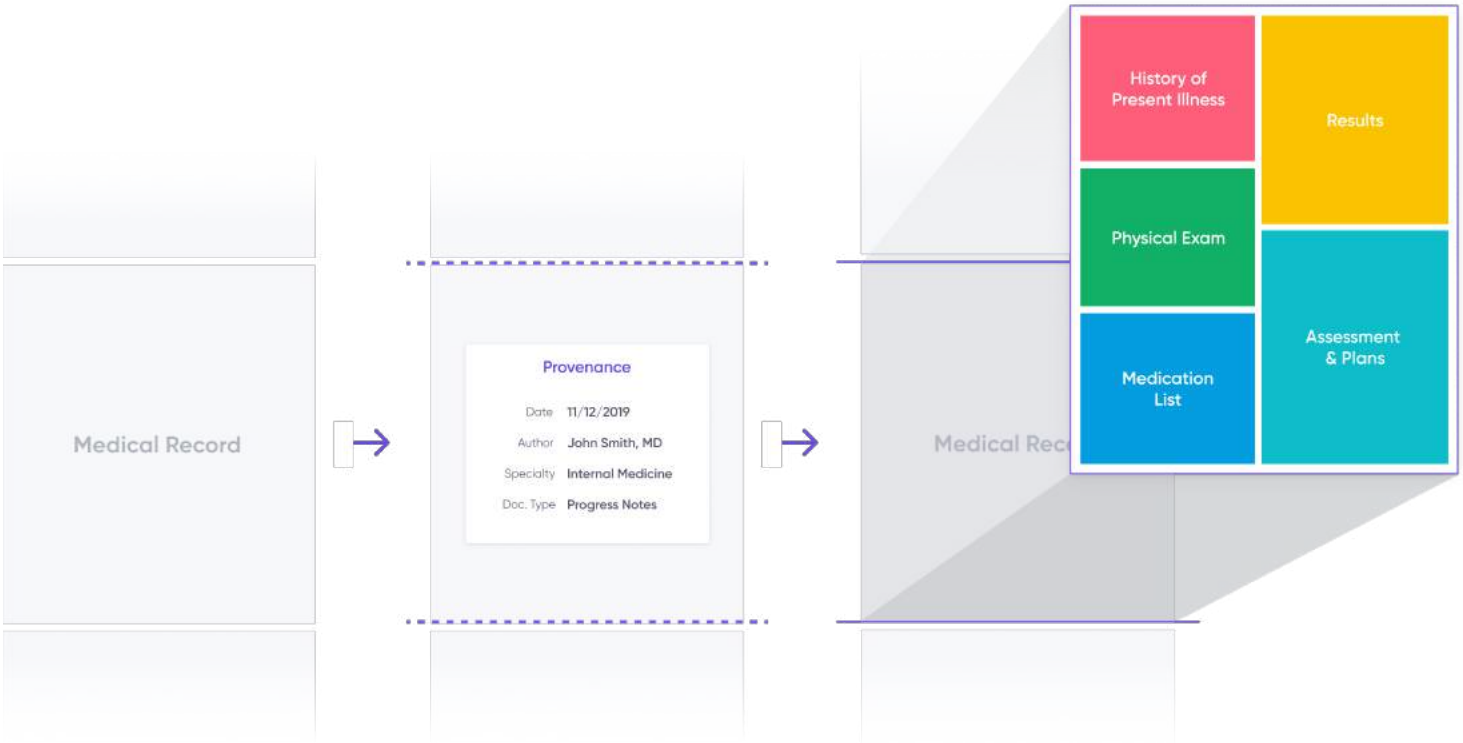
Figure 3 - To date, PicnicHealth has collected over 3 million medical records from 90,000+ providers across the US.

Documentation Types and Sectioning

The ability to trace each data entity to its documentation type and section provides greater transparency for regulators and allows researchers to differentiate the value of each data point. A diagnosis of sickle cell anemia is likely more reliable when it originates from a hematologist's consult note than from a surgeon's procedure note, and a mention of amoxicillin from a patient's medication list has a different clinical implication than when it appears in the allergy section.

At PicnicHealth, each PDF file is broken down into individual clinical documents, including various types of provider notes, diagnostic reports, and non-clinical documents. Key metadata including document author, date of service, and facility are associated with each document. For example, provider notes include any notes from in-person or virtual visits with a clinical provider, such as an outpatient progress note, an emergency department note, or a hospital admissions history and physical. Diagnostic reports include labs, pathology, genetics reports, etc. Other clinical documents may include email correspondences, facesheets and flowsheets. Non-clinical documents are items such as fax covers, standard disclosure and consent forms.

All clinical documents are then further divided into sections that enable targeted data extraction. For example, an outpatient progress note may include sections such as chief complaint, history of present illness, review of systems, problems list, social history, family history, medications, allergies, physical exam, assessment and plans.

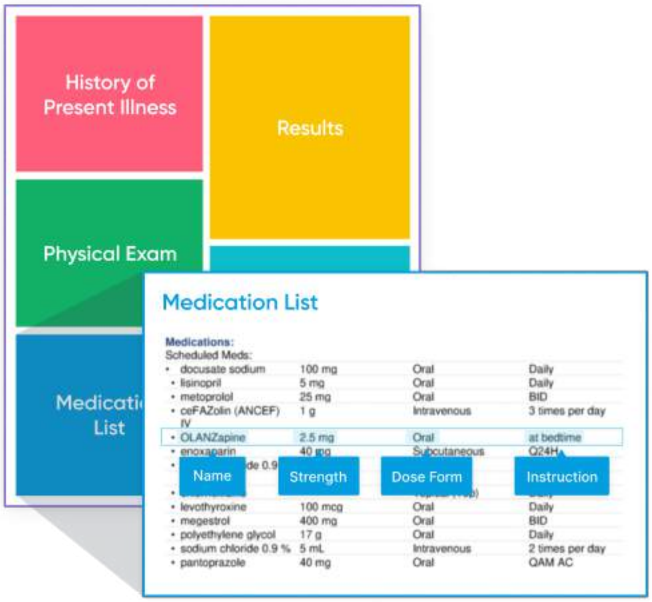


Date 11/12/2019
Author John Smith, MD
Specialty Internal Medicine
Doc. Type Progress Notes

Figure 4 - PDF files of medical records are segmented into individual documents and assigned document types. Each document is further divided into relevant sections to enable targeted data extraction.

Structured Data Extraction

Clinical concepts from structured portions of the document such as ICD-9/ICD-10 coded problem lists, medication lists, vital signs, and labs are identified and extracted from the record into a common format, and then mapped to standard ontologies (SNOMED, LOINC, RxNorm). Both the clinical concept itself (i..e. Olanzapine medication) and associated fields (i.e. medication strength, dose form, and instruction) are extracted. Every data point can be traced back to the original medical record through detailed metadata containing date, time, provider name, visit type, document type, etc.



Medication List

Medications:
Scheduled Meds:

• docusate sodium	100 mg	Oral	Daily
• lisinopril	5 mg	Oral	Daily
• metoprolol	25 mg	Oral	BID
• ceFAZolin (ANCEF) IV	1 g	Intravenous	3 times per day
• OLANZapine	2.5 mg	Oral	at bedtime
• enoxaparin	40 mg	Subcutaneous	Q24H
• levothyroxine	100 mcg	Oral	Daily
• megestrol	400 mg	Oral	BID
• polyethylene glycol	17 g	Oral	Daily
• sodium chloride 0.9 %	5 mL	Intravenous	2 times per day
• pantoprazole	40 mg	Oral	QAM AC

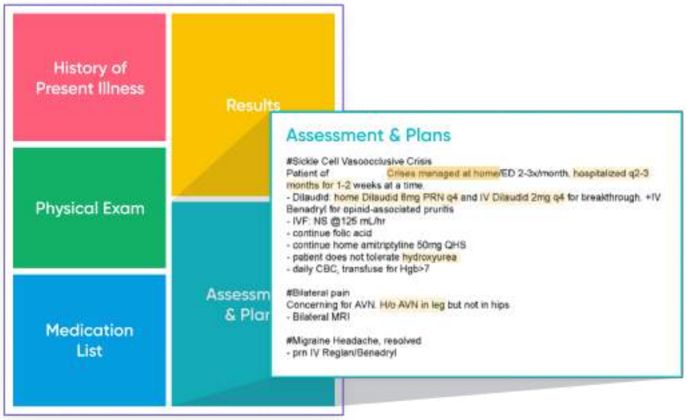
Figure 5 - The PicnicHealth machine learning model can predict all essential properties for a given medication (name, dose, unit, etc.)

PicnicHealth uses proprietary machine learning and AI models to improve the identification and extraction process from free text, for example using concept embedding based on a deep neural language model trained on our dataset of records. The primary role of machine learning at PicnicHealth is to improve the efficiency of record processing at scale, enabling us to cost-effectively process large volumes of medical information without sacrificing the accuracy provided by trained medical professionals. Every piece of data extracted from a record is checked by two trained human annotators, one to confirm the accuracy of the machine learning model and a second for quality assurance. Regular internal audits yield 96%-99% inter and intra-rater agreement.

Narrative Text Data Extraction

Data is also extracted from unstructured narrative portions of the text to answer key research questions. For clinical concepts embedded in narrative text, PicnicHealth works with research partners and disease-area experts to define disease-specific variables and outcomes of interest. For any clinical concept, the concept itself, as well as any related fields and metadata surrounding the clinical concept are captured. Examples of data that are only available through narrative data abstraction include disease severity status such as cancer staging, patient symptoms, tumor size, disease progression and response to treatments, and other physician assessments.

Regular expressions (regex) are used to predict each clinical concept. The sensitivity and specificity of models are tested and refined based on feedback from human reviewers. As with standard data elements, all concepts are mapped to common medical ontologies, associated fields and metadata are captured, and all final data extracted are reviewed by two separate trained annotators. Unlike stand-alone NLP algorithms, PicnicHealth's combined ML and human approach is able to accurately capture contextual information such as timing (past, current, future), certainty (yes, no, possible), and disease progression information (worsening, recurrence, metastasis).



Assessment & Plans

#Sickle Cell Vasoocclusive Crisis
Patient of Crises managed at home/ED 2-3x/month, hospitalized q2-3 months for 1-2 weeks at a time.
- Dilaudid: home Dilaudid 8mg PRN q4 and IV Dilaudid 2mg q4 for breakthrough. +IV Benadryl for opioid-associated pruritis
- IVF: NS @125 mL/hr
- continue folic acid
- continue home amitriptyline 50mg QHS
- patient does not tolerate hydroxyurea
- daily CBC, transfuse for Hgb>7

#Bilateral pain
Concerning for AVN. H/o AVN in leg but not in hips
- Bilateral MRI

#Migraine Headache, resolved
- prn IV Reglan/Benadryl

Figure 6 - Clinical concepts can be extracted from narrative sections of medical records through machine predictions and human reviews.

DICOM images

DICOM images allow researchers to validate and further explore radiographic outcomes that are not readily available in text. ML-guided algorithms may be used to extract key clinical findings contained within the images.

All DICOM images sent to our facility are uploaded into the PicnicHealth platform and de-identified. Metadata is erased from every image, DICOM IDs are renamed to a new random ID, and any image that includes personally identifiable information (PHI) in the image itself is suppressed from exporting.

Summary

The PicnicHealth data extraction process is machine-guided and human-curated. Data is richer than standard EHR data as we are able to extract disease specific information from narrative portions of the medical records in addition to extracting structured data. We adhere to the highest security and ethical standards in handling personally identifiable information, and patients remain in control of their data throughout the process.

Appendix

The following is a list of potentially identifiable data elements that are modified or removed prior to sharing data with partners.

Data Element	Process
Patient first, middle, and last name	Never transmitted
Patient SSN	Never transmitted
Patient address	Never transmitted
Patient phone number	Never transmitted
Patient email address, URL, social media handles	Never transmitted
Patient IP addresses or MAC addreses	Never transmitted
Patient web browser history/cookie data	Never transmitted
Patient website logins or passwords	Never transmitted
Patient birth date	Transmitted as YYYY-MM or YYYY only
Patient biometric identifiers	Never transmitted
Patient photographs	Never transmitted
Health insurance identifiers	Never transmitted
Medical facility information	Medical facility name and address are transmitted
Radiology DICOM images	All information including but not limited to: accession number; institution name/address; referring physician name, phone number, and address; station name; institutional department name; performing physician name; patient identifier; patient name, address, phone; patient date of birth; patient insurance identifiers; military rank; clinical trial identifiers are removed and not transmitted. Study ID, Series ID, patient ID, and SOP instance ID are removed and replaced with internal PH codes.