WHITEPAPER

Unstructured data: an untapped resource of patient data for clinical trials

Executive Summary

With the increased use of digital apps, services, and documentation, unstructured data (i.e., free-form clinical notes, imaging, pathology reports, patient-reported outcomes) now represent a large portion of the available patient information in healthcare. As an abundant source of insights into patient care and outcomes, this data source has become the target of many initiatives across the healthcare and life sciences industries. These include clinical data registries aimed at understanding diseases and improving the quality of patient care, such as the National Institutes of Health's Alzheimer's Prevention Registry or the Rare Diseases Registry Program (RaDaR); chronic disease management; understanding family history; research using real-world patient data; and improving recruitment strategies for clinical trials.

The challenge with harnessing the value of unstructured data sources lies in their diversity and disparate locations — as well as changes in systems and data standards over time. Over the past couple of decades, there have been many attempts at creating a robust technological solution to overcome these roadblocks. We're at an exciting time where technology, and our understanding of the data, has advanced to a point at which efficiently and effectively mining unstructured data is a possibility. Innovative approaches using the best-fit solution for the specific challenge at hand will help the industry leverage all its data sources to improve disease understanding, treatment development and patient care and outcomes.

The amount of digital patient data across healthcare settings has grown exponentially over the past decade. Electronic health records (EHRs) have become standard in the majority of healthcare organizations,[1] driven in part by the need to document patient care to qualify for performance-based reimbursement from the United States Centers for Medicare and Medicaid Services (CMS). These digital records have improved patient-provider communication, allowed easy patient access to doctor's notes and test results, and enabled online prescription management. Beyond patient care, they are a valuable source of information for clinical trials and research using real-world patient data.

Yet, extracting data in a meaningful way for these purposes remains challenging. While data entry for some fields is standardized to specific terminologies (and therefore relatively easy to access and analyze), a large portion of patient information is captured in free-form clinical notes, laboratory reports, and imaging findings — also known as unstructured data. Extracting the wealth of insight from these unstructured data is not straightforward. Technological solutions have successfully addressed some of the challenges with accessing detailed information about the patient journey from both structured and unstructured data. Future enhancements will continue to fill in the gaps.

Structured data provide certain insights

Structured data include those that are coded according to a standardized format, such as LOINC,[2] ICD-10 (soon to be ICD-11),[3] and SNOMED.[4] Developed to help with the exchange of health data between systems, these standards provide clear rules on what data to enter and how to enter those data. Generally, their development was driven by a specific need, such as reimbursement or to order lab tests.

Structured fields include those for patient demographics (age, gender, height, weight), vital signs (blood pressure, heart rate), some laboratory tests, and medications. Because the data already exist in a fixed structure, analysis is fairly straightforward. Moreover, mappings between standards have also been developed to facilitate transferring data from one standard to another.

Unstructured data capture nuanced patient information

The majority of data in EHRs are unstructured: typed and written text, photos, radiological images, pathology slides, video, audio, streaming device data, PDF files and faxes. In addition, data that were structured in an older format can also be considered unstructured once it is no longer supported.

These data sources include nuanced information about the patient's disease state, clinical sequence, or specific treatments that is difficult or unnecessary to capture in codes. In a recent study, unstructured data (i.e., ECG images and clinical notes) were required to identify 96.4% of patient encounters for acute coronary syndrome (ACS), and only 0.1% of encounters that met the inclusion criteria were identified solely based on diagnostic codes.[5] ICD-9 and current procedure terminology (CPT) codes identified <11% of cases with adverse events related to central venous catheters in another study using EHR data.[6]

Unstructured data were required to identify

96.4%

of patient encounters for acute coronary syndrome (ACS)

AND ONLY

0.1%

of encounters were identified solely based on diagnostic codes.

The use of coded fields is often driven by reimbursement. Therefore, if somwething being documented isn't billable, it is unlikely to be included in structured data. For example, details about the care episode or hospital stay are often captured in free-text discharge summaries. New phenomena, such as COVID-19, lack codes initially. Until a code is created, documentation relies on unstructured notes.

In addition, some information might only be represented textually because it is not a clinical indicator for a disease/disorder in its own right. This could happen for diseases/conditions that have a combination of symptoms that, on their own, are not specific. Normal pressure hydrocephalus is one such disorder: excess cerebrospinal fluid in the brain causes thinking and reasoning problems, difficulty walking, and loss of bladder control. Because it typically affects older adults and the symptoms happen irregularly and separately, they can easily be attributed to age, and the symptoms themselves might be coded (e.g., a fall). Diagnosis of these types of disorders is challenging, particularly when different episodes over time are documented throughout the patient's records across multiple systems.

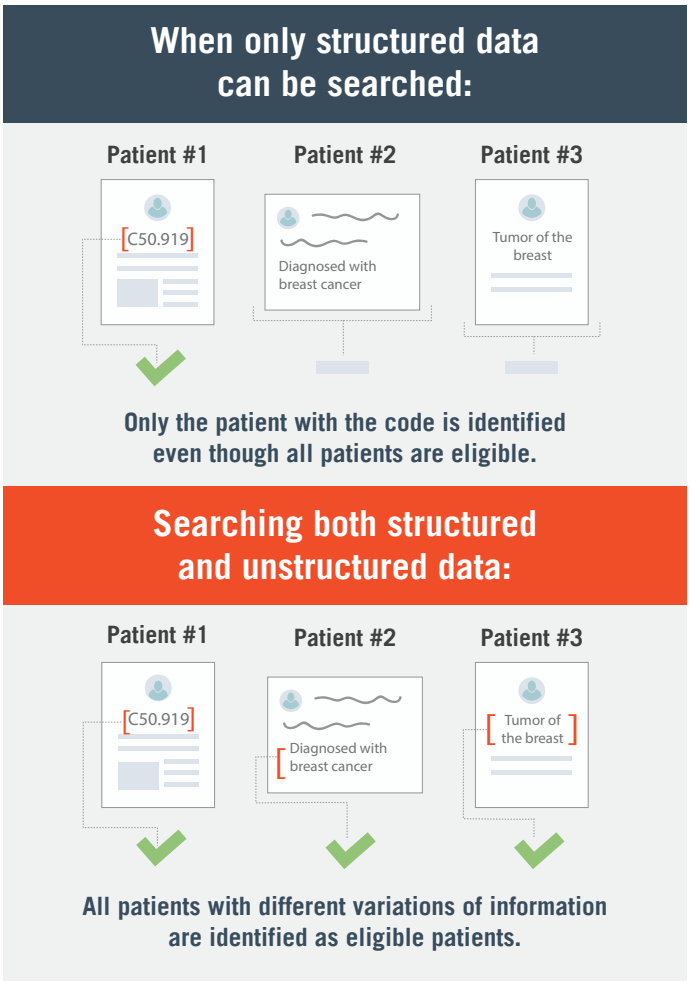
EHR data can expedite research

EHR data have the potential to accelerate the research lifecycle and get treatments to patients faster, by supporting protocol feasibility assessments, site assessments, patient recruitment, real-world research, and synthetic cohorts.

Using health system data, sponsors can evaluate whether there are sufficient patient numbers to meet the protocol inclusion/exclusion criteria. If not and the criteria are too limiting, the protocol can be modified accordingly — before the trial starts. Similarly, once the study design is finalized, specific clinical sites with potentially eligible patients can be identified. Designing a study for the best chances of recruitment and determining where patients are located can minimize the risk of slow or stalled recruitment and enrollment. During the trial, eligible patients can be identified and recruited based on their EHR data.

Alternative research designs are also possible using EHR data, such as research using real-world data. Synthetic cohorts based on standard of care, as documented in the EHR, also reduce the need for recruitment of a comparison group.

To achieve these goals, researchers are attempting to use systems and standards that were developed for other purposes such as documentation of care or communication between healthcare systems, which remains a limitation. To find relevant information for patient identification, manual searches of EHR data are typically necessary. This can be accomplished with structured data relatively easily, but searching unstructured data is time-consuming, and data are easily missed.



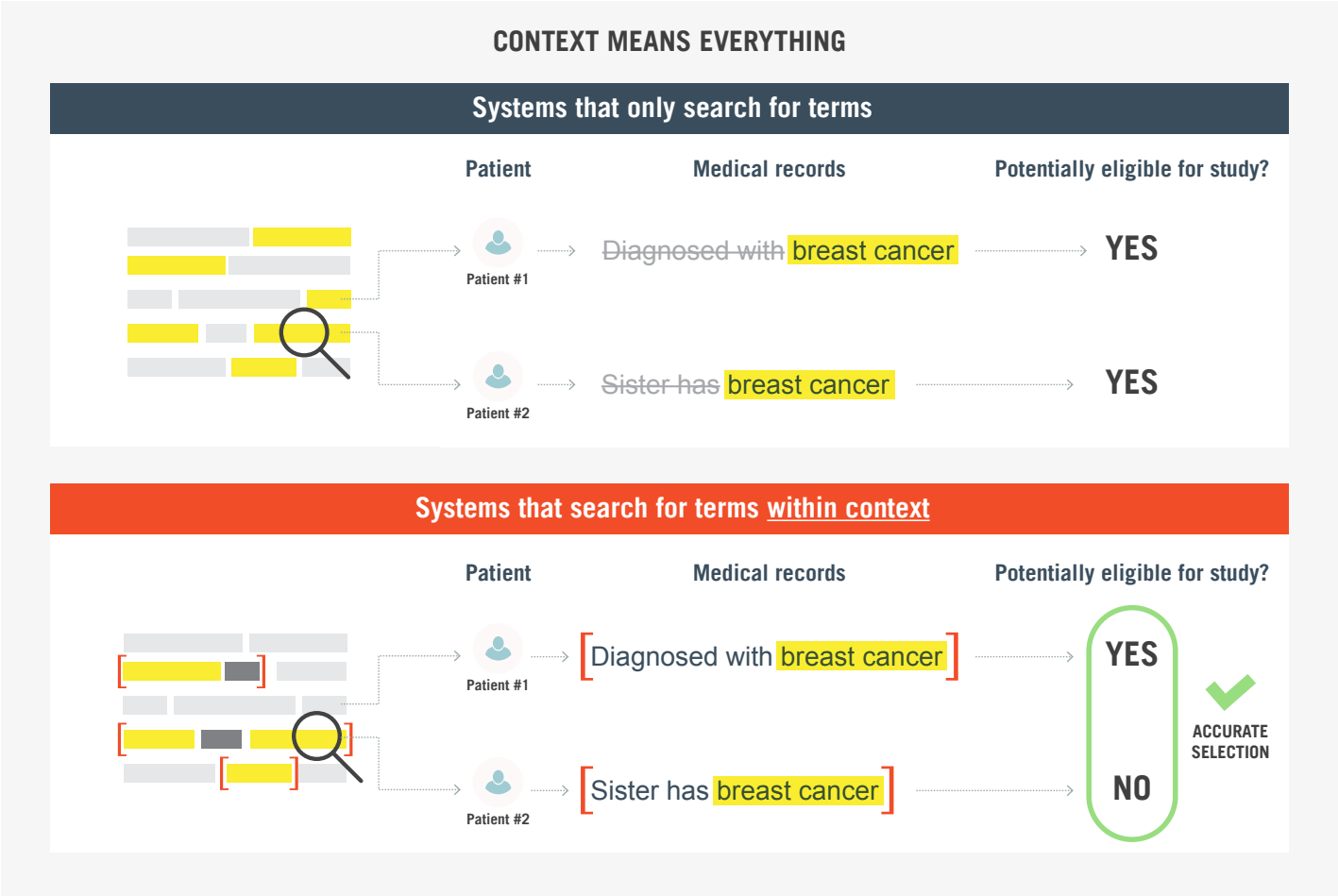
Unstructured data are crucial for research but difficult to analyze

Given the rich patient information and additional insights in unstructured data sources, they are essential to obtaining a holistic patient overview and being confident that a criteria-based search returns an accurate list of patients.[7]

However, due to their inherent nature, unstructured data sources need to be extracted, processed, and normalized in order to be analyzed. Complicating this process are grammatical and spelling errors, ambiguities, differences in terminology within and across institutions, and varying abbreviations that make healthcare data uniquely noisy.[8-10]

Complex medical terms are often abbreviated, but not always consistently, and the same abbreviation can be used for different terms (e.g., ER for emergency room and estrogen receptor; MG/mg for myasthenia gravis and milligrams). Abbreviations and acronyms can change over time as our understanding of a disease or disorder improves or with modified terminology: chronic obstructive pulmonary disorder (COPD) historically was, and still can be, called chronic obstructive lung disease (COLD).

Therefore, context is important. Searching for COLD (chronic obstructive lung disease) could also return results for the common cold. Similarly, abbreviations that are commonly used terms can vary in their meaning based on the word order: AIDS blindness vs blindness aids. In a colonoscopy report, "bleeding" has different meanings depending on whether it is found in the indications, findings, or recommendations section.



Another consideration of the context is the specific terms that are being searched and whether they change over time. For example, for studies including social determinants of health (SDH) in their criteria, it might be pertinent to only search the last 30 days to reflect current status.

While it would be understandable to assume that scientific and medical language is consistently used across physicians, disciplines, and sites, it is actually very unspecific and difficult to automatically parse and extract.

Technology solutions increase the usability of unstructured data

Technological solutions that aim to reliably and accurately extract both structured and unstructured data from EHRs for research purposes have been developed, with varying success.[6,11,12] These systems rely on natural language processing (NLP) and machine learning techniques, among others.

Natural language processing (NLP)

NLP attempts to bridge the gap between natural human language and computer understanding using an amalgamation of artificial intelligence (AI), linguistics, and computer science. NLP methods include text classification, entity recognition (manually developed resources such as dictionaries, regular expressions; machine learning), language modelling, word embeddings, coreference resolution, negation detection, and positive-only labeling.[13]

These methods gather and process the data to ensure accuracy, completeness, and consistency: tokenization (splitting sentences into words), lemmatization (converting single words into their base form), and parts-of-speech tagging (identifying words as nouns, adjectives, verbs, etc). It is also necessary to understand the syntactical features of human language, such as the correct order of terms and morphological properties (e.g., prefix, suffix, infix). Finally, the words need to be assigned a meaning.

Challenges for NLP include the disambiguation of abbreviations, the corpus assembly, and differing document structures, EHR systems, local policies and practices, and documentation practices by healthcare providers.[8] Each requires a different strategy, which necessitates a more complex implementation.

However, NLP-based methods have been successfully used to extract patient information. High performance of NLP with EHR data has been documented for colonoscopy quality measures, as measured by accuracy, recall, precision, and F measure.[8] To identify the occurrence of falls, multiple NLP methods were used on clinical notes, resulting in an average precision of 0.590 to 0.764.[14] A clinical text mining system consisting of several NLP approaches recently also successfully identified fall history, including temporal fall patterns, with high recall.[15] Moderate performance for NLP to infer SDH has also been reported, which the authors attribute to the lexical and semantic ambiguity associated specifically with SDH.[10]

Machine learning

Machine learning and deep learning algorithms can be used to address some of the challenges with NLP. These algorithms extract features from data and learn from datasets in order to become proficient at specific tasks, such as recognizing relevant terms and patterns in EHR data. They include support vector machines (SVMs), decision trees, convolutional neural networks (CNNs), long short-term memory (LSTM), and generative adversarial networks.

SVMs have been successfully used to extract medical record information for heart disease, diabetes, and breast radiology.[13] Similarly, heart disease, multiple sclerosis, and oncology data have been detected by naïve Bayes, while random forests have been used to predict heart disease, classify cancer types, and identify hypertension.[13] Neural networks in general perform well at extracting relevant patterns from

sequential data, while CNNs are effective at imaging analysis (MRIs, CTs, etc), which fills a gap associated with NLP.

Graph processing of ontologies

An important step in deriving meaning from healthcare data is understanding the relationships between the terms found in those data. Ontologies are data models that formally describe the concepts and relationships between them. They enable the sharing of knowledge. Typically, ontologies are described as having classes, relationships, and attributes. These can be visually depicted in graph-based models and used to draw conclusions about the similarities between terms and concepts.

To effectively depict healthcare information, which varies considerably across discipline, by timepoint, and geographically, multiple ontologies from many sources (e.g., academic, guidelines, local use) should be graphed and continuously refreshed as new ontologies are created or existing ontologies are updated.

Language models

Beyond the published ontologies, context around that information provides valuable information. Language modelling uses statistical and probabilistic methods to analyze text in NLP applications. The text is fed through algorithms that have the established rules for the context for the task at hand. In healthcare, these rules can include which stop words are important and whether capitalization plays a role in meaning.

Equally important is the ability of the language model to process words that might derive meaning from other words or terms that are in distant locations within the overall text. It needs to be able to understand references to those terms. Therefore, to be successful, language models require domain-specific knowledge and expertise.

Concept maps

While ontologies can depict specific terms and their relationships from defined ontologies, the determinants and presentations of health and ill health are complex. To capture all the nuances of a particular disease/disorder, disease state, treatment trajectory, and outcomes, concept maps are particularly useful. A concept map is a visual representation of knowledge (concepts and ideas) and its relationships. The resulting framework provides a basis for critical analysis of data.

They are typically arranged with the most general concepts at the highest level and more specific concepts arranged below. This structure depends on the domain and the context in which the knowledge is being considered. In healthcare, academic publications of relationships about diseases/disorders, medications, other treatments, and disease course are good sources, as are insights from subject matter experts and clinician experience. New and changing knowledge and the impact on these relationships should also be folded in iteratively.

COVID: The need for data before it was a recognized entity

To extract data, that data must first be understood. With a new situation such as COVID-19, there is a need to understand what is occurring before terms have officially been defined, and definitely before they are widely recognized and used.

In March 2020, when the pandemic was still in its infancy, COVID-19 was being documented in medical records in many ways. However, by understanding how it was being documented at partner hospitals, the Deep 6 AI team could quickly add those terms to the corpus and incorporate it into the concept maps. This enabled patients with suspected or diagnosed COVID-19 to be identified before there was coding language for it.

Flexibility is key

Given the many, long-standing challenges of extracting and understanding unstructured data and the advantages/disadvantages of each technological approach, it is important to consider hybrid approaches and not rely on a single technology. This is particularly true as the solutions evolve and improve; maintain the flexibility to adopt and adapt the approach that will provide the most accurate results for the end user.

The combination of solutions might need to allow for greater recall than precision. Low recall will likely miss cases, resulting in human time to read the clinical notes to determine if some were missed. On the other hand, low precision will return more potentially ineligible cases, but it takes less time (when the system allows it) to review the data source to eliminate those cases.

After all, the net outcome of implementing a system to search EHR data for eligible patients should be a reduction in time spent validating the patient list via a manual review of patient records.

What is the future?

As is true across many industries, AI technologies for healthcare data are evolving. Research to find efficient ways to analyze unstructured data (and extract meaning) is happening at a faster pace. AI has moved from rules-based systems to neural networks. The former require rigorous definition of rules and vocabularies, and it is difficult to include all possibilities. Neural networks have demonstrated superior performance in detecting patterns and deep relationships.[13] They also have the benefit of requiring less data to make logical decisions, enabling storage of smaller amounts of data.

Methods are also being refined for patient-level inference of disease state and progression. Analysis of temporal relationships of data and linking different events or entries within a statistical space could enable identification and diagnosis of diseases such as normal pressure hydrocephalus. Storing information as dimensional data for a particular patient instead of as data points is one approach.

Temporal relation extraction can truly build a patient's story. This applies not only to progression but also the sequence of events, treatments, and response. In oncology, second-line therapy is often used when the patient doesn't respond to first-line therapy. Being able to easily detect previous therapies and the patient's response to those would help with patient selection. Previous studies have evaluated several methods for temporal extraction, again with varying success.[13,15]

Conclusion

The development of methodologies on EHRs has enabled the power of those data to be extracted. As an industry, we are moving closer to being able to enable researchers and clinicians to search for a concept or term that fits within the semantic space in which the trial protocol was written — and receive an accurate result. Future progress will build on these developments and expand our capabilities to quickly and efficiently identify patients meeting the most granular of criteria.

References

1. Adoption of Electronic Health Record Systems among U.S. Non-Federal Acute Care Hospitals: 2008-2015. May 2016. Available at https://www.healthit.gov/sites/default/files/briefs/2015_hospital_adoption_db_v17.pdf. Accessed on August 20, 2021.
2. LOINC. Available at https://loinc.org/. Accessed on August 20, 2021.
3. International Statistical Classification of Diseases and Related Health Problems (ICD). Available at https://www.who.int/standards/classifications/classification-of-diseases. Accessed on August 20, 2021.
4. SNOMED International. https://www.snomed.org/. Accessed on August 20, 2021.
5. Tam CS, Gullick J, Saavedra A et al. Combining structured and unstructured data in EMRs to create clinically-defined EMR-derived cohorts. BMC Med Inform Decis Mak 2021;21:91. https://doi.org/10.1186/s12911-021-01441-w
6. Penz JF, Wilcox AB, Hurdle JF. Automated identification of adverse events related to central venous catheters. J Biomed Inform 2007;40(2):174-82.
7. Wei WQ, Teixeira PL, Mo H, et al. Combining billing codes, clinical notes, and medications from electronic health records provides superior phenotyping performance. J Am Med Inform Assoc. 2016;23(e1):e20-e27. doi:10.1093/jamia/ocv130
8. Carrell DS, Schoen RE, Leffler DA, et al. Challenges in adapting existing clinical natural language processing systems to multiple, diverse health care settings. Journal of the American Medical Informatics Association 2017:24(5):986-991. https://doi.org/10.1093/jamia/ocx039
9. Jensen K, Soguero-Ruiz C, Mikalsen KO, et al. Analysis of free text in electronic health records for identification of cancer patient trajectories. Scientific Reports 2017;7:46226. https://doi.org/10.1038/srep46226
10. Feller DJ, Bear Don't Walk Iv OJ, Zucker J, et al. Detecting Social and Behavioral Determinants of Health with Structured and Free-Text Clinical Data. Appl Clin Inform 2020;11(1):172-181. https://doi.org/10.1055/s-0040-1702214
11. Edgcomb J, Zima B. Machine learning, natural language processing, and the electronic health record: Innovations in mental health services research. Psychiatric Services 2009;70:346-349.
12. Ehrentraut C, Ekholm M, Tanushi H, et al. Detecting hospital-acquired infections: A document classification approach using support vector machines and gradient tree boosting. Health Informatics Journal 2016;24:24-42.
13. Sheikhalishahi S, Miotto R, Dudley JT, et al. Natural Language Processing of Clinical Notes on Chronic Diseases: Systematic Review. JMIR Med Inform. 2019;7(2):e12239. doi:10.2196/12239
14. Korach ZT, Yang J, Rossetti SC, et al. Mining clinical phrases from nursing notes to discover risk factors of patient deterioration. Int J Med Inform. 2020;135:104053. doi:10.1016/j.ijmedinf.2019.104053
15. Topaz M, Murga L, Gaddis KM, McDonald MV, et al. Mining fall-related information in clinical notes: Comparison of rule-based and novel word embedding-based machine learning approaches. Journal of Biomedical Informatics 2019;90:103103. https://doi.org/10.1016/j.jbi.2019.103103.