INVITED COMMENTARY

A Health Care Clinical Data Platform for Rapid Deployment of Artificial Intelligence and Machine Learning Algorithms for Cancer Care and Oncology Clinical Trials

Soma Sengupta, Rohan Rao, Zachary Kaufman, Timothy J. Stuhlmiller, Kenny K. Wong, Santosh Kesari, Mark A. Shapiro, Glenn A. Kramer

The xCures platform aggregates, organizes, structures, and normalizes clinical EMR data across care sites, utilizing advanced technologies for near real-time access. The platform generates data in a format to support clinical care, accelerate research, and promote artificial intelligence/machine learning algorithm development, highlighted by a clinical decision support algorithm for precision oncology.

Introduction

Artificial intelligence (AI), which includes natural language processing (NLP) and large language models (LLM) as well as machine learning (ML), is increasingly being applied to electronic medical records (EMR) to enhance health care delivery and research. These technologies facilitate the querying, understanding, and extraction of information from EMRs, which is often unstructured and voluminous.

AI-based methods, including ML classifiers that automatically categorize data, have been effectively applied to query EMR, demonstrating high diagnostic accuracy across multiple organ systems [1]. NLP, particularly when combined with ML and deep learning, can be used to automatically analyze medical documents, extracting information and classifying it into predefined categories [2]. This can improve health care delivery by aiding clinicians in finding relevant information or gaining insights from medical reports and clinical research forms.

Named entity recognition (NER) models, such as ClinSpacy and MedSpacy, have been developed for clinical NLP to identify and extract specific medical concepts such as drug names and dosages from clinic notes, which is crucial for converting unstructured data into a structured format for downstream analysis [3]. Similarly, ML techniques have been applied to extract real-world data variables from unstructured EMR, such as information on cancer diagnoses and treatments, which can be used for evidence generation in oncology [4].

EMRs contain valuable medical information but are often difficult for computers to understand due to jargon and abbreviations. Since Google's release of the Transformer architecture, LLMs have entered mainstream use [5]. Following the public release of the LLM known as GPT-3, many organizations have generated models based on similar architecture to create bespoke LLMs trained on health care data to extract and interpret EMR, enabling more personalized clinical recommendations [6–7].

To make AI/ML solutions useful in health care and fulfill their promise of improving patient care, easing clinician workloads, and accelerating clinical research, several challenges remain [8]. First, there is a shortage of available EMR training data, due to data privacy and security concerns. Second, there are complex issues with training models to understand the mix of structured and unstructured data in the records. Third, there is great diversity in the data models underlying EMR systems. Finally, the cost of training a custom LLM means that hospitals and health systems should generally not attempt to create their own. Thus, a general purpose clinical LLM that could integrate with all or most existing health care IT is yet to be realized. Here, we describe a decentralized precision oncology platform for gathering, organizing, structuring, and standardizing longitudinal clinical data directly from EMRs for use in downstream AI and ML applications.

Methods

xCures Platform for Medical Record Aggregation and Data Structuring

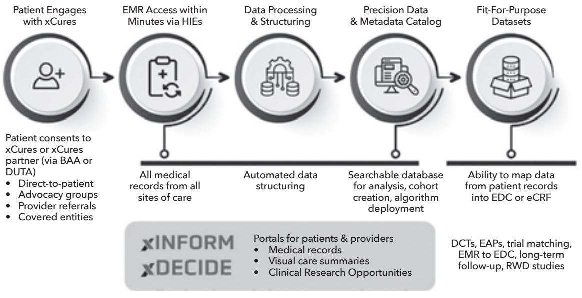
The xCures platform (summarized in Figure 1) provides an infrastructure that allows for near real-time clinical data

Electronically published July 10, 2024.

Address correspondence to Soma Sengupta, Bioinformatics 5115, UNC School of Medicine, Chapel Hill, NC 27599 (ssengup@email.unc.edu) and Glenn A. Kramer (gkramer@xcures.com).

N C Med J. 2024;85(4):270-273. ©2024 by the North Carolina Institute of Medicine and The Duke Endowment. All rights reserved. 0029-2559/2024/85415

FIGURE 1.
Overview of the xCures Platform


Patient Engages with xCures
EMR Access within Minutes via HIEs
Data Processing & Structuring
Precision Data & Metadata Catalog
Fit-For-Purpose Datasets
Patient consents to xCures or xCures partner (via BAA or DUTA)
• Direct-to-patient
• Advocacy groups
• Provider referrals
• Covered entities
All medical records from all sites of care
Automated data structuring
Searchable database for analysis, cohort creation, algorithm deployment
Ability to map data from patient records into EDC or eCRF
xINFORM
xDECIDE
Portals for patients & providers
• Medical records
• Visual care summaries
• Clinical Research Opportunities
DCTs, EAPs, trial matching, EMR to EDC, long-term follow-up, RWD studies

Patients are enrolled on the platform via direct consent or through partnership with provider systems or covered entities. Electronic medical records (EMR) from diverse sources are aggregated in near-real time via health information exchanges (HIEs), classified, and structured using NER, NLP, and LLM models, yielding a standardized data lake and metadata catalog suitable for cohort creation, data analysis, and algorithm development. The platform offers an annotation layer to generate fit-for-purpose datasets in support of decentralized trials (DCTs), expanded access protocols (EAPs), and real-world data (RWD) studies among other uses. Personal web portals for patients (xINFORM) and providers (xDECIDE) are available to view aggregated medical records, clinical summaries, and support research.

aggregation, structuring, and outcomes tracking across sites of care. The platform operates entirely in HIPAA-compliant cloud infrastructure, relieving burden on local servers at institutions. The platform is available to cancer patients and their providers at health care institutions to support clinical decision-making, such as second opinions and molecular tumor boards (groups of experts with different specialties who meet to discuss patient treatment options). The collected data are utilized in an institutional review board (IRB)-approved registry (XCELSIOR: NCT03793088) that includes development of AI/ML tools to improve the collection of data for use in research, including development of AI systems.

The xCures platform utilizes four technological innovations to permit rapid, near real-time access to comprehensive structured clinical data from all sites of care:

1. Connectivity with local, state, and national health information networks. Within 15 minutes, the platform can gather (or refresh) raw exports from a wide array of EMR systems (e.g., EPIC, CERNER, OncoEMR) in Consolidated Clinical Document Architecture (C-CDA) and standardize that data across EMR data models into a standardized Fast Healthcare Interoperability Resources (FHIR) format. This process includes conversion of scanned PDFs and image files (JPEG, TIFF, etc.) through optical character recognition (OCR) and natural language processing.
2. An OCR and document classification algorithm helps to classify and organize each page of PDF (or other image) files in the medical records received.
3. An NER pipeline helps to identify clinical concepts and their synonyms in the raw text of every medical document and has the ability to search for these terms across medical records within the platform.
4. NLP, ML, and LLM algorithms then auto-structure data elements from text including clinically important features such as cancer stage/grade, histology/morphology, tumor body location, medications, procedures, patient race/ethnicity, performance scores (KPS/ ECOG), and next-generation sequencing (NGS) results from commercial labs (e.g., Foundation Medicine, Caris, Tempus, Guardant).

The result is a platform with more complete patient data than is found in any one EMR with information easily searchable and sortable after conversion into a common data model. This makes it easy for clinicians to understand the full treatment history, medical history, pathology and “omics” data (generated from technology used to study the genome, transcriptome, metabolome, and other “omes” of an organism), labs, vitals, imaging results, and other information. This is foundational and enables generation of graphical or timeline views, such as a Gantt chart-style patient journey that can rapidly convey key information for more efficient clinical decisions.

The platform allows medical records to be sent directly via fax, through connections with laboratories, and via patient-directed approaches such as file uploads or sharing from patient portals. The platform accepts medical data in many different formats, ranging from structured electronic formats such as FHIR and C-CDA to electronic PDFs, to uploaded images and other text documents that are processed via OCR and classified by document type. The platform can integrate data into a single system from multiple EMRs that typically exist at large practices (cancer center, inpatient hospital, outpatient clinic, infusion clinics, pharmacies, etc.) in ways that often do not exist in medical practice.

All medical records are digitized and housed on cloud systems suitable for data abstraction and queries of free text or raw terms across all patients in the database. This facilitates algorithm development by institutional data scientists.

Clinical Data Standardization and Data Annotation for Downstream Applications

Data are auto-structured to a FHIR R4 data model and coded to OMOP-based ontologies, and molecular biomarkers and genomic findings are standardized. Each data element is coded and standardized as part of the data extraction process. Data elements that are extracted include conditions/comorbidities, medications, procedures, labs, vitals, and biomarkers including genomic alterations from next-generation sequencing.

The platform also has integrated annotation capabilities to both validate auto-extracted data and capture custom data elements that may not exist in standard dictionaries. All data are housed in a central database with an audit trail and provenance back to the source documents for ease of source data verification. Thus, the system is robust and streamlined for researchers and staff to utilize for population-level research or clinical trials.

Real-World Applications for Clinical Decision Support, Trial Matching, and Research

The xCures platform includes web portals for patients, providers, and institutional partners that permit electronic enrollment and viewing of medical records and structured clinical data. This facilitates near real-time clinical data access and sharing across all parties. For institutional partners, data across patients are aggregated to create a catalog of real-world data available to support clinical care and secondary research.

Decentralized applications. Patient (xINFORM) and provider (xDECIDE) web portals enable any cancer patient in the country to enroll for xCures services directly. Patients receive a personal health record and access to structured summaries auto generated by the xCures platform from comprehensive EMR data. Individual providers can view medical records for their patients and have them screened for treatment options including clinical trials and expanded access programs.

As a proof-of-concept clinical decision support application of the xCures platform, we designed a ML algorithm (xCORE: xCures Option Ranking Engine) utilizing structured data elements, outcomes data, and results from virtual tumor board reviews to surface potential treatment options for individual patients based on clinical features [9]. Clinical data elements including genomic alterations were extracted from EMR, validated by human annotators, standardized to discrete clinical features to create a "patient case vector," and processed by xCORE. The xCORE algorithm contains multiple submodels, which integrate to provide a probabilistic score for each of thousands of potential treatment options: 1) an NLP submodel trained on clinicaltrials.gov inclusion/exclusion criteria, 2) a biomarker-drug submodel that utilizes CiViC and DGIdb drug and gene databases, 3) a virtual tumor board consensus option ranking submodel, 4) an NLP submodel that extracts treatment information from rationale statements from virtual tumor boards and leading oncologists, and 5) a submodel that utilizes actual outcomes (overall survival, time-on-treatment, real world progression-free survival) from propensity score-matched advanced cancer patients enrolled in the XCELSIOR research protocol. The patient case vector is then independently distributed across each of the submodels and integrated to provide a probabilistic score of matching [9]. Initial testing of the algorithm shows the same sensitivity of this algorithm to inter-tumor board recommendations but with greater precision and F1 score (mean of precision and recall). As more patients are enrolled, the clinical utility of these recommendations is collected both directly from patients and providers and in the longitudinal outcomes generated by the system.

Institutional applications. The xCures platform is available to institutions directly as a standalone cloud software to enable a range of activities for personnel involved in informatics and information technology (CIO, IT Director), clinical research (e.g., CMO, Director of Clinical Research, CMIO, research coordinators) and clinical care (individual treating oncologists and nurses).

The platform provides near real-time access to EMR from all sites of care nationally and summarized medical histories, supporting providers in patient care, particularly for new patient visits. The data annotation layer permits researchers to produce and reference regulatory-compliant, source-verifiable data in standard, normalized, and structured formats that enable it to function as a common baseline to layer on diverse applications and analyses. The platform provides customizable dashboards of standardized data to characterize patient populations and compare the effectiveness of the treatment strategies employed in their care. Structured data elements and unstructured text are available in a format for rapid, programmatic screening of clinical trial inclusion and exclusion criteria with simple IF, THEN, AND, OR, NOT functions. The richness of clinical data available across all sites of care enables the development and deployment of comprehensive AI/ML algorithms.

Conclusion

The application of AI, NLP, NER, ML, and LLMs to EMR holds immense potential for revolutionizing health care delivery and clinical research. These technologies have demonstrated their effectiveness in querying, understanding, and extracting valuable information from the often unstructured and voluminous data within EMR. ML classifiers, NLP algorithms, and NER models have shown high diagnostic accuracy and efficiency in extracting specific medical concepts from EMR, facilitating improved health care delivery and clinical decision-making. However, the widespread implementation of these technologies requires addressing challenges such as data standardization and privacy concerns. As these technologies continue to evolve and integrate into health care systems, they hold the promise of enhancing patient care, accelerating clinical research, and ushering in a new era of precision medicine. NCMJ

Soma Sengupta, MD, PhD, MBA, FRCP (UK), FAAN, FANA Clinical Professor of Neurology, Division Chief of Neuro-Oncology, Co-Director of Neurofibromatosis Program, Vice-Chair of Research, Department of Neurosurgery, School of Medicine, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina.

Rohan Rao, MD Neurology Resident, Ronald Reagan UCLA Medical Center, University of California, Los Angeles, California.

Zachary Kaufman Vice President of Product Management, xCures, Oakland, California.

Timothy J. Stuhlmiller, PhD Vice President of Scientific and Medical Affairs, xCures, Oakland, California.

Kenny K. Wong, CGC Chief Product Officer, xCures, Oakland, California.

Santosh Kesari, MD, PhD Chair and Professor, Department of Translational Neurosciences, Saint John's Cancer Institute at Providence Saint John's Health Center, Santa Monica, California.

Mark A. Shapiro, MBA Chief Operating Officer, xCures, Oakland, California.

Glenn A. Kramer, PhD Chief Technology Officer, xCures, Oakland, California.

Acknowledgments

Soma Sengupta and Rohan Rao have no conflicts of interest to report. All other authors have employment and equity interest in xCures.

References

1. Liang H, Tsui BY, Ni H, et al. Evaluation and accurate diagnoses of pediatric diseases using artificial intelligence. *Nat Med.* 2019;25:433–438. https://doi.org/10.1038/s41591-018-0335-9
2. Crema C, Attardi G, Sartiano D, Redolfi A. Natural language processing in clinical neuroscience and psychiatry: a review. *Front Psychiatry.* 2022;13:946387. doi: 10.3389/fpsyt.2022.946387
3. Singh K. M*L4LHS/clinspacy: Clinical Natural Language Processing using spaCy, scispacy, and medspacy.* GitHub. 2021. Accessed May 1, 2024. https://github.com/ML4LHS/clinspacy
4. Adamson B, Waskom M, Blarre A, et al. Approach to machine learning for extraction of real-world data variables from electronic health records. *Front Pharmacol.* 2023;14:1180962. doi: 10.3389/fphar.2023.1180962
5. Vaswani A, Shazeer N, Parmar N, et al. Attention Is All You Need. June 12, 2017. arXiv:1706.03762
6. Agrawal M, Hegselmann S, Lang H, Kim Y, Sontag D. Large *Language Models are Few-Shot Clinical Information Extractors.* Empirical Methods in Natural Language Processing (EMNLP). 2023;142:102573. https://aclanthology.org/2022.emnlp-main.130.pdf
7. Yang X, Chen A, PourNejatian N, et al. A large language model for electronic health records. *npj Digit. Med.* 2022;5:194. https://doi.org/10.1038/s41746-022-00742-2
8. Wornow M, Xu Y, Thapa R, et al. The shaky foundations of large language models and foundation models for electronic health records. *npj Digit. Med.* 2023;6:135. https://doi.org/10.1038/s41746-023-00879-8
9. Shapiro MA, Stuhlmiller TJ, Wasserman Asher, et al. AI-augmented clinical decision support in a patient-centric precision oncology registry. *AI in Precision Oncology.* 2023;1:27-37. doi: 10.1089/aipo.2023.0001.