OFFICE OF THE SECRETARY
PATIENT-CENTERED OUTCOMES RESEARCH TRUST FUND

PROJECT REPORT

FINAL REPORT

Approaches to Developing Artificial Intelligence Training Datasets and Areas for Future Research

Prepared for
The Office of the Assistant Secretary for Planning and Evaluation (ASPE)
at the U.S. Department of Health and Human Services

by
NORC at the University of Chicago

July 2026

Office Of The Assistant Secretary For Planning And Evaluation

The Assistant Secretary for Planning and Evaluation (ASPE) advises the Secretary of the U.S. Department of Health and Human Services (HHS) on policy development in health, disability, human services, data, and science; and provides advice and analysis on economic policy. ASPE leads special initiatives; coordinates the Department's evaluation, research, and demonstration activities; and manages cross-Department planning activities such as strategic planning, legislative planning, and review of regulations. Integral to this role, ASPE conducts research and evaluation studies; develops policy analyses; and estimates the cost and benefits of policy alternatives under consideration by the Department or Congress.

Office Of The Secretary – Patient-Centered Outcomes Research Trust Fund

The Office of the Secretary Patient-Centered Outcomes Research Trust Fund (OS-PCORTF) was established as part of the 2010 Patient Protection and Affordable Care Act and is charged to build data capacity for patient-centered outcomes research. Coordinated by ASPE on behalf of the Department, OS-PCORTF has funded a rich portfolio of projects to meet emerging U.S. Department of Health and Human Services policy priorities and fill gaps in data infrastructure to enhance capabilities to collect, link, and analyze data for patient-centered outcomes research. For more information, visit https://aspe.hhs.gov/collaborations-committees-advisory-groups/os-pcortf.

This report was funded by the Office of the Secretary Patient-Centered Outcomes Research Trust Fund (OS-PCORTF) under Contract Number HHSP233201500048II of the HHS Office of the Assistant Secretary for Planning and Evaluation (ASPE). The work was carried out by NORC at the University of Chicago and ASPE. The authors are solely responsible for this document's contents, findings, and conclusions, which do not necessarily represent the views of HHS, ASPE, or NORC. Readers should not interpret any statement in this product as an official position of ASPE or of HHS.

Suggested Citation: Heaney-Huls K, Desai PJ, Aronoff A, Couture S, Kho A, Dullabh P. Approaches to Developing Artificial Intelligence Training Datasets and Areas for Future Research. Washington, D.C.: Office of the Assistant Secretary for Planning and Evaluation, U.S. Department of Health and Human Services. July 2026.

Contributing Authors

Krysta Heaney-Huls, MPH, Research Scientist, NORC
Priyanka J. Desai, PhD, Principal Research Scientist, NORC
Abigail Aronoff, MPH, Research Associate II, NORC
Sara Couture, MPH, Social Science Analyst, ASPE
Abel Kho, MD, Internist, Professor of Medicine and Preventive Medicine, Director of the Institute for Artificial Intelligence in Medicine, Northwestern University
Prashila Dullabh, MD, Vice President and Senior Fellow, NORC

Project Officers and Project Leadership

Sara Couture (ASPE)
Jennifer Wiltz (ASPE)
Jesse Anderson (ASPE)
Marcos Trevino (ASPE)
Prashila Dullabh (NORC)
Mithuna Srinivasan (NORC)

Table of Contents

Executive Summary **1**

Key Findings **1**

1. Introduction **4**

1.1 Objectives and Structures of the Report 4

2. Methods **5**

3. Findings **5**

3.1 The Landscape for AI Training Datasets Using Real or Synthetic EHR Data 5

3.2 Real-World AI Training Dataset Development 7

3.3 Synthetic AI Training Dataset Development 10

3.4 Approaches for Developing AI Training Datasets 12

3.5 Resources and Tools to Support AI Training Dataset Development and AI Model Development 14

4. Discussion **16**

5. Conclusion **17**

References **18**

Appendix A. Methods **24**

Appendix B. Publicly Available Training Datasets Identified **27**

Table of Figures

Exhibit ES-1. Real-world AI Dataset Development Steps and Associated Challenges **1**

Exhibit ES-2. Approaches for Improving AI Training Datasets **2**

Exhibit 1. Key Dataset & Access Characteristics at a Glance **6**

Exhibit 2. AI Model Development Case Studies **7**

Exhibit 3. AI Training Dataset Development **8**

Exhibit 4. Example De-identification Tools and Tool Accessibility **14**

Exhibit 5. Example NLP Tool and Tool Accessibility **15**

Exhibit A1. Literature Search Terms **24**

Exhibit A2. Relevant OS-PCORTF Projects **24**

Exhibit A3. Literature Search Inclusion and Exclusion Criteria **25**

Exhibit A4. Fields Included in Full Article Abstraction Form **25**

Executive Summary

The Office of the Secretary Patient-Centered Outcomes Research Trust Fund (OS-PCORTF) supports efforts to strengthen data infrastructure for patient-centered comparative clinical effectiveness research (CER), including the use of artificial intelligence (AI) to generate meaningful clinical insights that can directly impact treatments and care.

Electronic health record (EHR) data are a core research data source for CER. However, clinical notes and other free-text documents within the EHR remain underutilized due to their unstructured nature. Natural language processing (NLP) and other AI techniques can enable insights from unstructured clinical text from EHR. To unlock its potential, researchers and AI developers need access to gold standard AI training datasets that are representative, well-annotated, and validated. Despite their importance, such datasets are limited and often narrowly representative.

Key Findings

This report provides a summary of available AI training datasets and outlines guidance and resources to support improved dataset development and future research.

AI Training Dataset Landscape: Gold standard AI training datasets are characterized by rigorous, transparent, and iterative development, involving clinical domain experts for accurate annotation and validation, predefined quality criteria (e.g., completeness, consistency, representativeness), standardized annotation protocols and robust validation techniques, and feedback loops for continuous improvement. However, the majority of AI models used in research and practice are trained on proprietary training datasets, limiting transparency, reproducibility, and generalizability. As a result, most existing publicly available AI training datasets fall short of gold standard principles, creating barriers to model trustworthiness and comparability.

This report identifies 27 publicly available AI training datasets. Example uses of these AI training datasets include event classification, temporal reasoning, and predictive modeling. Approximately two-thirds of these AI training datasets represent data derived from real-world clinical records, mostly from acute care settings, about half of which are subsets of each other. The remaining datasets are synthetic. Less than half of the identified datasets provide documentation of the methodology used to create the AI training dataset or the annotation methods.

Real-World AI Training Dataset Development: Developing high-quality, real-world training datasets for AI is a complex, resource-intensive, multi-step process that requires precision and clinical expertise. This report describes five general steps: 1) defining the objective, 2) data collection and sampling, 3) data preprocessing, 4) data transformation, and 5) data annotation. Each step in the development process presents unique challenges (Exhibit ES-1). We also highlight rapid advances in AI that are transforming dataset development methods. AI tools are increasing used to improve the efficiency and accuracy of preprocessing, transformation, and annotation.

Exhibit ES-1. Real-world AI Dataset Development Steps and Associated Challenges

Step in Real-World AI Dataset Development	Key Challenge
Defining the Objective: Specifying the AI model's intended use case guides decisions on data types, labeling strategies, and scope	• Publicly available AI training datasets are insufficient to address researchers' needs. • Achieving a satisfactory balance between data quality, representativeness, annotation strategies, resource

Step in Real-World AI Dataset Development	Key Challenge
	constraints, and AI model performance.
Data Collection and Sampling: Identifying relevant clinical text and ensuring representative sampling	• Overly narrow sampling reduces data heterogeneity, which can impact AI performance.
Data Preprocessing: Cleaning, standardizing, and de-identifying free text data for extraction	• De-identification is labor-intensive, but necessary to minimize risk of protected health information exposure.
Data Transformation: Normalizing, or mapping, data to standard clinical terminologies (i.e., ontologies)	• Mapping to standard ontologies is complex. • Accurately identifying text for extraction is complicated by discontinuous mentions of clinical concepts and temporal relationships.
Data Annotation: Contextualizing extracted data using standardized schemas	• Manual annotation is costly and time-consuming, limiting scalability.

Synthetic AI Training Dataset Development: Synthetic data generation—using statistical models, deep learning, and large language models—has grown significantly. While synthetic data mitigates privacy risks and supports scalable model development, challenges remain around generalizability, fidelity, and the lack of standardized evaluation frameworks.

Approaches for Developing AI Training Datasets: Developing gold standard AI training datasets is essential for building accurate and reliable AI models. This report highlights strategies that can support the development of AI training datasets. Exhibit ES-2 summarizes strategies to improve AI training dataset development and downstream model performance.

Exhibit ES-2. Approaches for Improving AI Training Datasets

Strategy for Improving AI Training Dataset Development and Outputs	Implementation Guidance
 **Using varied data can improve model performance**	• Use data that vary by source, population characteristics, and clinical presentation • Supplement real-world data with synthetic data
 **Applying different AI data extraction techniques can enhance model precision and recall**	• Combine data that are already categorized (i.e., labeled) and data that are not (i.e., unlabeled) • Incorporate broader spans of discontinuous mentions (i.e., phrases where important clinical information is split up by unrelated text)
 **Using synthetic data in text extraction tasks can reduce risk of protected health information (PHI) exposure**	• Use synthetic data to improve how clinical concepts are recognized (i.e., named entity recognition) and automated de-identification
 **Using AI techniques can increase the efficiency of data annotation**	• Determine the correct order of clinical events before interpreting their clinical relevance to improve labeling • Leverage distant and weak supervision to annotate data • Leverage few-shot learning, a technique where AI learns a new task using a few examples instead of thousands, to annotate PHI
 **Integrating feedback loops can help to define clinical concepts and identify meaningful outcomes**	• Incorporate feedback loops to improve annotator guidelines

Strategy for Improving AI Training Dataset Development and Outputs	Implementation Guidance
Standardizing annotation protocols can enhance consistency	• Assess opportunities to refine annotator guidelines to improve reliability

Areas for Future Research: Strengthening data capacity in the following key areas can help overcome barriers to developing gold standard AI training datasets:

- Establish industry-wide metrics for de-identification and synthetic data quality.
- Expand datasets to include multimodal data and broader patient populations.
- Enhance annotation efficiency through participatory and community-based approaches.
- Develop benchmark datasets to assess privacy risks and model generalizability.

Addressing these data infrastructure needs will help accelerate the development of trustworthy AI systems that can improve patient health outcomes.

1. Introduction

The U.S. Department of Health and Human Services (HHS) Office of the Secretary Patient-Centered Outcomes Research Trust Fund (OS-PCORTF)[1] supports initiatives that strengthen data infrastructure for patient-centered outcomes research (PCOR). These efforts aim to advance the development and use of emerging technologies—such as artificial intelligence (AI)—to generate high-quality evidence on the effectiveness of treatments, services, and interventions that matter most to patients, caregivers, clinicians, and policymakers.[2]

Natural language processing (NLP) is an evolving AI technique that enables the extraction of meaningful insights from unstructured clinical text, such as electronic health record (EHR) notes.[3] Some of the most valuable information for patients, caregivers, and decision makers—particularly details that reflect patient experiences, clinical decision-making, and care delivery—is often embedded in free text.[4] However, because this information is unstructured and, until recently, not easily readable by machines, it remains largely underutilized in AI applications and patient-centered research.

Gold standard science is a set of research principles emphasizing the use of reproducible and transparent methods.[5] According to *America's AI Action Plan*, for AI models to accurately mine through unstructured free clinical texts across different applications, researchers must have access to gold standard AI training datasets that are representative, well-annotated, and ethically sourced.[6,7] High-quality AI training datasets are essential national assets that can accelerate progress toward gold standard medicine.[6] Still, only a limited number of datasets are publicly available, and many of these are narrowly representative because they are largely derived from specialty care settings and well-resourced academic institutions.[8]

This report and accompanying inventory provide a summary of the current landscape of publicly available AI training datasets and identify key technical resources to support dataset and AI model development. Drawing on approaches for enhancing the performance of AI models, the report offers guidance for researchers seeking to strengthen their own real-world and synthetic AI training development processes. The AI landscape is advancing at an extraordinary pace, driven in large part by innovations such as large language models (LLMs), which enable sophisticated and adaptive AI approaches. This report describes strategies to overcome limitations in the current AI training dataset infrastructure to strengthen research and AI training dataset development, and highlights how rapid progress in AI is transforming approaches to dataset creation.

1.1 Objectives and Structures of the Report

This report describes the findings of a literature review identifying best practices, key considerations, and current gaps for creating and using AI training datasets in practice. This project was guided by the following research questions:

1. What are the current best practices for developing gold standard AI training datasets used for research on health outcomes and effectiveness of interventions built with EHR data?
2. What are current best practices for creating de-identified and synthetic AI training datasets for research on the outcomes and effectiveness of interventions used in health care?
3. What characteristics or tools are researchers looking for in AI training datasets outside of "best practices" (e.g., codebooks, guides)?
4. What AI training datasets built with EHR data are currently available to researchers? How accessible are those datasets? What are the critical gaps in the current infrastructure?

The sections of the report are organized as follows: the ***Methods*** section provides an overview of the literature review methods; the ***Findings*** section describes a) the current landscape of AI training datasets, b) strategies and challenges associated with real-world and synthetic AI training dataset development, c) approaches for developing gold standard AI datasets, and d) available resources for AI training dataset development; the ***Discussion*** section describes opportunities to strengthen AI dataset development; and the report ends with a ***Conclusion*** section that summarizes findings. The report also contains supplementary appendices described in the following sections.

2. Methods

The literature review was conducted to identify peer-reviewed and grey literature that details AI training dataset development and characteristics of publicly available datasets. Additionally, the review identified gaps and opportunities to strengthen the current data infrastructure for AI training dataset development.

To complete the review, we conducted targeted searches in PubMed and Google Scholar to identify peer-reviewed and grey literature describing AI training datasets derived from free clinical text within EHRs **(*Appendix A*, Exhibit A1)**. Additionally, we conducted targeted searches of six OS-PCORTF projects with relevant work related to AI training dataset development (***Appendix A*, Exhibit A2**). Articles identified through the literature review underwent a two-step screening process: an initial title and abstract review, followed by a full-text review for eligibility. We applied search terms and inclusion/exclusion criteria as detailed in ***Appendix A*, Exhibit A3.**

In total, we screened 114 peer-reviewed articles, 31 of which were included after full-text abstraction. Our initial search yielded 93 peer-reviewed articles, and we conducted targeted supplemental searches on de-identification techniques, synthetic datasets, and dataset creation methodologies, screening an additional 21 articles. We also screened 44 pieces of grey literature, 34 of which were included after full-text abstraction. Finally, we conducted targeted searches of related references to further inform the AI training dataset inventory described in the next section.

Information was abstracted into a data abstraction form (***Appendix A*, Exhibit A4**) that was accessible to all team members to ensure transparency and collaborative input. A designated team member (who did not perform the original abstraction) validated the accuracy and consistency of the abstractions. We conducted thematic content analysis of the literature findings to develop the report and accompanying inventory.[9]

3. Findings

From the literature review, we identified 26 publicly available AI training datasets **(*Appendix B*)** and synthesized key themes and opportunities for developing AI training datasets to support AI model development. In this section we describe these findings by detailing: 1) the landscape for AI training datasets using real or synthetic EHR data, 2) real-world AI training dataset development and best practices for creating gold standard datasets and de-identification, 3) synthetic AI training dataset development and best practices, and 4) resources to support AI training dataset development and AI model development and validation.

3.1 The Landscape for AI Training Datasets Using Real or Synthetic EHR Data

A "gold standard" AI training dataset serves as the foundation for developing reliable and high-performing AI models, particularly in clinical and patient-centered research. Drawing from principles outlined in the White House's *Gold Standard Science* initiative[5] and findings from our literature review, we framed a gold standard AI

training dataset as one developed through a rigorous, transparent, and iterative process. This process includes the active engagement of clinical domain experts, whose contextual knowledge is critical for accurate data annotation and validation. Their expertise ensures that datasets faithfully reflect clinical realities and support meaningful, patient-centered model development.

Beyond expert involvement, gold standard datasets are distinguished using predefined criteria to assess and validate data quality. These criteria commonly include measures of completeness, consistency, and representativeness across patient populations and care settings. The development process also incorporates mechanisms to evaluate methodological rigor, such as standardized annotation protocols and robust data validation techniques. Finally, gold standard dataset creation involves feedback loops that facilitate continuous improvement based on stakeholder input, performance evaluation, and emerging best practices.[10] This iterative approach helps ensure that the resulting datasets are not only technically robust but also ethically and clinically relevant.

To explore how current practices compare to gold standard principles, we reviewed 27 publicly available AI training datasets. While these resources provide valuable insights, it is important to acknowledge that many AI models in practice are trained on proprietary datasets that are not publicly accessible, often due to the high costs of data sharing and limited representativeness across diverse healthcare populations.[11] The accompanying inventory (**Appendix B**) details characteristics such as data type (real-world or synthetic), data sources, key variables, sample size, dataset creation methodologies, and example use cases.

Dataset Characteristics. Eighteen datasets are based on real-world data, shown in **Exhibit 1**. Within these datasets, 13 are original datasets and five are derived from the Medical Information Mart for Intensive Care (MIMIC) database,[12] and one from the Temporal Histories of Your Medical Event (THYME) corpus.[13] Most datasets (n=16) are derived from acute care settings; other care settings include critical care (n=9) and specialty care (n=7). Less than half of the datasets have documented dataset creation methodologies (n=10) and annotation methodologies (n=7). Of the synthetic datasets (n=9), six describe their synthetic data validation methodologies. Example use cases for each dataset include event classification, temporal reasoning, predictive modeling, population-level analyses, and model benchmarking.

Exhibit 1. Key Dataset & Access Characteristics at a Glance

Dataset Characteristics	Data Access Characteristics
18 real-world datasets (across acute, critical, and specialty care settings)	8 datasets available for immediate download
9 synthetic datasets	6 publicly available user guides
10 documented dataset creation methodologies and 7 annotation methodologies	9 datasets requiring DUA/IRB approval
6 described synthetic data validation methodologies	10 datasets requiring a contractual mechanism

Dataset Accessibility. Eight datasets are readily available for download **(Exhibit 1)**. The remaining datasets require a data use agreement (DUA) or Institutional Review Board (IRB) approval for access (n=9) or are available through contracting mechanisms (n=10). Additionally, six datasets have publicly available user documentation to support dataset implementation.

Dataset Use Case Studies. Of the above datasets, two of the most commonly leveraged sources include the MIMIC database[12] and Synthea synthetic patient generator.[14] These datasets have been widely used to support innovation in clinical research and AI model development (**Exhibit 2**).

Exhibit 2. AI Model Development Case Studies

MIMIC: De-identified Patient Data

MIMIC data are derived from 360,000 patients admitted to Beth Israel Deaconess Medical Center (BIDMC) between 2001 and 2022, and the Northwestern ICU database, a critical care dataset collected from a network of hospitals located in Chicago and harmonized with the MIMIC data structure. The Northwestern ICU database includes de-identified data from Northwestern Memorial HealthCare from 2020-2022.

Du et al. (2021)[15] developed IoHAN (Interpretable Outcome Prediction Model with Hierarchical Attention), a deep learning framework that predicts patient outcomes from electronic health record (EHR) data while providing clear, interpretable results. Unlike earlier models that treated EHR data as a single sequence, IoHAN recognizes the natural structure of records of individual hospital visits and the medical codes linked to each visit. It uses a layered attention system to identify which medical codes and visits most influence a patient's predicted outcome. This approach helps clinicians understand *why* the model made a certain prediction, improving trust and transparency in AI-driven healthcare.

IoHAN was trained and tested on the **MIMIC-III dataset** for mortality prediction, readmission risk, and disease risk. The model represented each patient's history as a series of visits, with the medical codes from each visit converted into numerical embeddings. In evaluation, IoHAN outperformed several state-of-the-art models in predictive accuracy and provided clearer data interpretations to influence care decisions.

Synthea: Synthetic Data

Synthea produces synthetic patient records for over one million simulated patients, modeled on real-world population health statistics and clinical workflows.[16]

Chen et al. (2022)[17] used **Synthea-generated** synthetic patient data to simulate a machine learning-enabled Learning Health System (LHS). They began with an initial cohort of 30,000 synthetic patients and sequentially added four additional batches of 30,000 patients each, resulting in a total of 150,000 patients, to mimic the ongoing accumulation of data in an LHS. Among these patients, more than 5,500 had lung cancer, forming a subgroup used to train and test risk-prediction models. The authors ran the same simulation for stroke risk prediction using the full synthetic cohort. Both tasks achieved high recall and area under the curve (AUC) scores, demonstrating the feasibility of leveraging synthetic data to prototype and evaluate LHS processes.

3.2 Real-World AI Training Dataset Development

Developing high-quality real-world AI training datasets for AI applications in health care is a complex, multi-step process that demands careful planning, technical precision, and clinical expertise—making it both resource-intensive and time-consuming.[18] Rapid advances in AI have fundamentally reshaped AI training dataset development methods. This section outlines the key stages of AI training dataset development, gleaned from studies published in the last decade, illustrated in **Exhibit 3**. This section highlights common challenges encountered in some of the steps, such as a lack of external validation and a lack of generalizability across clinical settings due to reliance on proprietary data or a limited set of widely used annotations and shared tasks.[19] We also highlight how recent advancements in AI are transforming training dataset development; for example, advances in AI tools are changing how data preparation steps, such as preprocessing, transformation, and annotation, are performed. While these tools can reduce manual effort and introduce new efficiencies, the steps themselves remain essential for ensuring data quality, interoperability, and clinical relevance.

Exhibit 3. AI Training Dataset Development

1. Defining the Objective

The first step in creating an AI training dataset is to clearly define the intended use case for the AI model. This involves specifying what the model is expected to learn or predict, which in turn guides decisions about the types of data needed and how that data should be labeled. When publicly available datasets are insufficient, researchers may need to build their own, tailoring the dataset to the specific clinical or research question at hand. Because building AI training datasets is a complex process, it is important for researchers to balance considerations such as data quality, representativeness, annotation strategies, and potential limitations, such as resource constraints. Alternative approaches, such as augmenting existing datasets, using synthetic data, or incorporating multiple sources, may also be necessary to ensure that the dataset adequately supports the model's objectives.

2. Data Collection and Sampling

Data collection begins with identifying relevant clinical text and corpora (collections of text), such as EHR notes, discharge summaries, and pathology reports. Sampling strategies must be carefully designed to ensure that the dataset captures a representative range of clinical scenarios. Key considerations include selecting appropriate document types (e.g., outpatient versus inpatient notes), metadata (e.g., patient demographics), and timeframes (e.g., pre- and post-treatment).[20]

Overly narrow or highly specific sampling criteria, especially for clinical concepts that use similar or overlapping terminology, can limit the diversity of the dataset and hinder the model's ability to recognize rare conditions or atypical presentations. For example, Sivarajkumar et al. (2024)[21] encountered challenges in extracting sleep-related information from clinical notes because concepts such as *snoring* and *sleep apnea* often overlapped but were inconsistently annotated, complicating downstream analysis.

3. Data Preprocessing

Preprocessing transforms raw clinical text into a structured format suitable for annotation and model training. This step includes cleaning the data, standardizing formats, and de-identifying protected health information (PHI).[22] Techniques such as document deduplication, keyword extraction, and named entity recognition (NER) are commonly used.[21,23,24,25]

De-identification is a critical step in preparing real-world clinical text for use in AI model development. It involves removing or obfuscating PHI from clinical documentation. De-identification methods must comply with Health Insurance Portability and Accountability Act (HIPAA) standards, using either Safe Harbor or Expert Determination methods detailed below: [26,27]

- **Safe Harbor:** This approach requires the removal of 18 specific identifiers (e.g., names, geographic details smaller than a state, dates directly related to an individual, contact information, and biometric identifiers).

- **Expert Determination:** This method involves a qualified expert applying statistical or scientific principles to determine that the risk of re-identification is very small. This approach allows for more flexibility and can preserve greater data utility, but it requires rigorous documentation and validation.

In practice, a range of techniques—including rule-based systems, AI models, and large language models (LLMs)—are used to detect and remove PHI. One common approach is clinical NER, which identifies and classifies sensitive information embedded in free-text clinical notes.[27,28,29]

Despite technological advances, de-identification presents several challenges. Manual de-identification is labor-intensive and costly, while automated methods—such as rule-based or machine learning approaches—often struggle to generalize across diverse data types and require large volumes of labeled examples to be effective.[27,30]

Beyond these operational challenges, newer, large-scale models trained on sensitive health data may introduce additional risks. Even with direct identifiers removed, these models can inadvertently retain or reveal PHI. Further, techniques like reverse engineering, in which trained models are manipulated to reconstruct sensitive information, raise significant privacy concerns.[31]

To mitigate these risks, privacy-preserving frameworks like differential privacy can be employed. Differential privacy works by injecting carefully calibrated statistical noise during model training or output generation, ensuring that the presence or absence of any individual's data does not significantly influence the results.[23,32] This guarantees that analyses remain effectively indistinguishable regardless of whether a specific data point is included, thereby minimizing the risk of re-identification.[33] Additionally, real PHI can be replaced with statistically representative synthetic data to reduce leakage risk.[34]

4. Data Transformation

Historically, after the clinical text was cleaned and de-identified, the next step was to transform the data into a format suitable for AI model training. This process typically involves normalizing the data and mapping clinical concepts to standardized medical ontologies, such as SNOMED-CT.[20] Doing so ensures consistency across datasets and enhances interoperability, making the data more compatible with downstream applications in machine learning and clinical research. However, data transformation can also alter the original context and introduce noise.[35] Therefore, a key decision during this phase is determining the appropriate level of granularity for coded concepts. Researchers must align the depth and specificity of selected codes with the intended clinical or research use case to ensure relevance and utility.

Recent advances in AI have introduced important shifts in how data transformation is approached. Today, LLMs can process raw text directly using embedding-based approaches, reducing the need for extensive manual transformation.[36] As a result, converting data into semi-structured formats may not always be necessary for initial dataset generation. However, data transformation will still play a role later in the pipeline to align outputs with standardized terminologies, ensuring interoperability and consistency across systems.

Data transformation presents several challenges. For instance, accurately extracting clinical concepts using NER can be difficult when dealing with discontinuous entity mentions (i.e., where relevant terms are interrupted by unrelated words).[37] Another common issue is determining temporal relationships, such as the sequence and timing of clinical events, which can be especially complex when EHR data contain multiple or conflicting time references. Addressing these challenges is essential to ensure the transformed data supports robust and clinically meaningful AI model development.[38]

5. Data Annotation

Annotation is a critical phase where clinical domain experts label the data using standardized classification codes.[20] Data annotation provides richer, more nuanced context to AI training data. Because of the complexity, data annotation has traditionally relied on domain experts to accurately contextualize the data. This human component adds time and costs to AI training dataset development. Yet, the better and more consistent the annotations, the better AI models learn. Zhu et al. (2024)[39] outline the critical steps of data annotation:

- **Annotator Training.** Novice annotators are paired with domain experts to learn the annotation purpose and clinical context, gradually building confidence through guided review and expert feedback.
- **Vocabulary Identification.** A coding dictionary is created using standardized ontologies or expert input to define the clinical concepts and terms that will be annotated.
- **Annotation Schema Development.** Annotators collaboratively develop and refine a standardized schema through iterative annotation and consensus-building to ensure consistency and completeness.
- **Annotation Execution.** Annotators apply the finalized schema to label clinical text, typically using structured formats like spreadsheets to record the presence or absence of specific features.
- **Result Validation.** A sample of annotated data are compared to expert-reviewed ground truth using metrics like precision, recall, and interrater reliability to assess and improve annotation quality.

Annotation can be unsupervised (i.e., identify entities without labeled data), semi-supervised (i.e., a small set of labeled data with a larger pool of unlabeled data), or automated, depending on the complexity of the task and available resources.[18,39] Manual annotation requires domain expertise, introducing substantial cost to AI training dataset development.[40] High-quality annotation is essential for creating reliable training AI datasets and often requires substantial investment in expert time and training.

3.3 Synthetic AI Training Dataset Development

Synthetic data generation offers a scalable solution for increasing the volume of data for AI model training, thus promoting AI model robustness. Overall, synthetic data helps overcome key limitations of real-world datasets, supporting safer and faster development of AI systems. Unlike real-world data, synthetic data mitigates privacy concerns—particularly around PHI—by reducing re-identification risks. It also streamlines the data creation process; for example, generating synthetic clinical notes has been shown to be more efficient than manual processes.

There are three main categories of methods used to create synthetic AI training datasets: 1) advanced statistical modeling and machine learning (such as those used to create the Synthea training datasets),[10] 2) deep learning models (i.e., neural networks), and 3) LLMs. Commercial LLMs such as ChatGPT Llama are increasingly used to generate synthetic data. For example, a team of researchers used GPT 3.5 and Llama to create synthetic patient summaries that were then used to compare the performance of an LLM in classifying acute renal failure from clinical notes.[25] These approaches have proven efficient and effective in generating synthetic notes. [23,25,41,42]

Once generated, AI developers validate the synthetic data to ensure the data reflects the structure, semantics, and variability of real-world clinical text. Validation involves comparing synthetic outputs against annotated reference datasets or expert-reviewed samples to assess fidelity and consistency. Multiple methodologies exist for validation, including Generative Adversarial Network-test[43] and cross-validation.[23,44]

Challenges to Using Synthetic AI Training Datasets for AI Development. The use of synthetic data for health care and research applications has expanded significantly in the past five years, with studies demonstrating the successful use of synthetic data across a range of clinical applications (e.g., phenotype classification, de-identification).[25,45] Alongside this growth, researchers have increasingly focused on validating the quality and utility of these synthetic datasets for training AI. Limitations to the use of synthetic data center on generalizability and privacy.

When synthetic notes are not representative of the outcomes across different patient demographics, model learning and performance are impacted.[41] Conversely, synthetic data based on small sample sizes or generated from data that contains biases will replicate these patterns when used to train AI algorithms.[46] To mitigate these risks, basic oversampling techniques, such as Synthetic Minority Over-sampling TEchnique (SMOTE), can rebalance datasets by increasing representation of underrepresented data.[47] More advanced methods, such as Generative Adversarial Networks (GAN) go further by producing realistic synthetic data through an iterative process between a generator and a discriminator. The generator aims to create data that mimic the real distribution while the discriminator evaluates whether the data are real or synthetic, driving both models to improve iteratively until the generated samples become indistinguishable from real-world data. GAN-based approaches have demonstrated better performance over traditional oversampling, particularly for complex tasks such as time-series modeling, while also supporting patient privacy.[48]

Another limitation of synthetic data is the complexity of the data or use case. One study found that Synthea and other synthetic patient generators do not fully account for how care might be delivered in real life (e.g., deviations from clinical guidelines) and thus have difficulty modeling the results of those real-world variations, such as non-compliance with clinical guidelines and geographic differences in health care utilization.[49,50] In this study, researchers evaluated Synthea's synthetic patient data by comparing the rates of four quality measures (e.g., colorectal cancer screening, controlling high blood pressure) against a representative sample of real-world patient data. The researchers found that while Synthea was fairly reliable in modeling patient demographics and the probabilities of health care utilization, it was limited in its capabilities to model a range of post-intervention health outcomes. New synthetic AI training datasets representing more clinical domains and multi-modal data (e.g., imaging data, wearable technologies) could help address this gap by simulating more real-world evolving clinical scenarios. However, there is heterogeneity in the methods and metrics used to evaluate the quality of synthetic data, making it difficult for researchers and AI developers to generalize potential AI performance across disease domains.[45]

Using LLMs to generate synthetic clinical data comes with specific challenges. One concern is the risk that LLMs may fail to exclude all PHI (e.g., location details) in the data,[23] as well as the potential for re-identification.[33] Differential privacy-based methods to generate synthetic data can reduce this risk by introducing a controlled quantity of noise in synthetic data, maintaining the privacy of personal information.[33] Another challenge is that LLMs require immense computing power, which can limit their accessibility and scalability. One solution is deploying optimized versions of LLMs on edge devices (i.e., hardware near data sources), which can reduce costs, improve privacy, and lower latency. However, these models typically rely on compression techniques and have narrower capabilities compared to full-scale models.[51,52] Finally, using LLMs can lead to model collapse. The risk of model collapse can occur when an AI model trained on LLM generated synthetic data repeatedly reinforces its own biases, creating a negative feedback loop that leads to increasingly biased or low-quality results. To address this, researchers can use bias detection and fairness auditing tools, such as IBM AI Fairness 360,[53] Microsoft Fairlearn,[54] and EthicalXAI Bias Engine,[55] which apply fairness metrics to identify and mitigate bias in AI systems.

3.4 Approaches for Developing AI Training Datasets

Training AI models on gold standard datasets supports the development of models with greater accuracy and reliability in performing core AI functions such as classification, pattern recognition, and outcome prediction. Despite the critical role of training datasets in AI development, there is limited literature detailing the processes involved in their creation, making it difficult to identify best practices. Given this gap, we examined AI model development studies to identify strategies researchers and AI developers have employed to enhance model performance. From this analysis, we identified several approaches that can support the development of gold standard AI training datasets—characterized by high-quality annotation and data with broader generalizability. The approaches described below align with three of the five AI training dataset development steps: 1) data collection and sampling, 2) data preprocessing, and 3) data annotation.

Data Collection and Sampling. Using data that vary by source, population characteristics, and clinical presentation in AI training datasets can help to balance precision and recall in AI models.

- **Integrate synthetic and real-world data**. Data heterogeneity can be achieved by supplementing AI training datasets derived from real-world data with synthetic notes to improve model performance.[25,44] This approach also has the potential to mitigate the challenge of overfitting.
- **Data Preprocessing.** Clinical text extraction involves identifying relevant text from unstructured EHR clinical notes. More flexible rules for identifying text can improve AI model precision and recall. Additionally, using synthetic data in text extraction tasks can reduce the risk of PHI exposure.
- **Use labeled and unlabeled data.** Using both labeled and unlabeled data can help boost data extraction performance, particularly for categories of data where limited information is available.[18,29] For example, researchers demonstrated that using an entity-specific, rather than a fixed, confidence threshold can improve the performance of machine learning approaches that use both labeled and unlabeled data in model training (i.e., semi-supervised learning). Using different thresholds across all clinical concepts used to recognize and extract prescription information from discharge summaries improved the quality of the AI training data. This approach also enhances the model's capability to recognize meaningful context that might not be captured in a dictionary, as well as identify medication mentions that are limited.[18] Semi-supervised learning is a useful approach in situations where manual labeling is too expensive or time-consuming.
- **Incorporate broader spans of discontinuous mentions.** Including broader spans of discontinuous mentions in the AI training dataset can help capture a more accurate representation of the clinical phenomena.[56]
- **Use synthetic data to improve NER and automated de-identification**. Synthetic data offers a promising solution—not only as a substitute for real-world data, but also to enhance NER performance.[57] AI developers can also use synthetic clinical notes to locally train AI models to better recognize PHI in real-world clinical documents.[28,58] This approach avoids privacy risks of using commercial LLMs to generate synthetic data derived from real-world data and reduces reliance on large volumes of manually labeled data. This approach can also support improvements in fine-tuning by optimizing synthetic data generation that is more representative of difficult-to-classify entities.
- **Use keyword extraction for synthetic note generation.** Another approach to enhancing patient privacy protection is using keyword extraction, the process of automatically identifying and extracting relevant text, in the development of synthetic clinical notes. One study found that the risk of PHI exposure was lower when using keyword extraction compared to one-shot generation (i.e., clinical notes generated from one prompt without iteration).[23]

Data Annotation. Data labeling and annotation are the processes of categorizing data to make the data usable for machine learning algorithms. It is also a critical part of supervised learning, where labeled data are used for training AI models. Several strategies have emerged to reduce reliance on manual annotation as well as improve the volume and quality of the annotations.

- **Use temporal sequencing**. Temporal sequencing, where annotators first determine the correct order of clinical events before interpreting their clinical relevance, can help reduce ambiguity during labeling.[59]
- **Leverage distant and weak supervision**. Distant supervision, the NLP technique of automatically generating large volumes of AI training data using structured rules or knowledge bases, minimizes reliance on manual annotation. [60] Additionally, leveraging weak supervision, an AI technique using noisy, incomplete, or imprecise labels, and multiple ontologies can support more robust annotation strategies. For example, training weak supervision models to recognize clinical concepts using existing medical ontologies and expert-generated rules can significantly reduce the need for manually labeled datasets, and enhance privacy protections by eliminating the need to share actual patient data.[40] Two promising directions include: (1) using unsupervised structured learning to enable multi-ontology labeling and (2) developing multi-agent or multi-layer models to enhance domain-specific concept recognition across clinical areas. This can help AI models better recognize, identify, and standardize clinical concepts.
- **Leverage few-shot learning**. Providing a few high-quality examples in a LLM prompt, a technique known as few-shot learning, can improve PHI annotation accuracy when using LLMs to generate synthetic clinical notes.[28]
- **Assess opportunities to refine annotator guidelines**. Two critical steps for achieving efficiency and quality of annotation are 1) training annotators and familiarizing them with the text corpus and 2) developing the annotation schema. Inter-annotator agreement, a measure of how consistently different annotators label the same data, can be used to assess the quality (i.e., consistency and reliability) of the annotations and the annotator guidelines. The inter-rater agreement score can indicate how well the annotation is defined.[56,59] A low score can signal opportunities to refine the annotator guidelines.
- **Incorporate patient and stakeholder involvement in the annotation process.** Patient and stakeholder involvement in the annotation process can enhance dataset quality and relevance. Patients offer insights into lived experiences, symptoms, and terminology that reflect real-world language and nuances, ensuring that labeling reflects patient-centered perspectives (e.g., how conditions impact daily life). Clinicians and care teams can validate clinical relevance and ensure annotations align with medical standards. Co-design, advisory groups, and iterative feedback loops are effective mechanisms for integrating stakeholder perspectives into annotation guidelines and validation steps. For example, one way to improve annotator guideline development is to incorporate feedback mechanisms where clinicians and patients can provide input on how clinical concepts are defined and interpreted.41 This participatory approach ensures that annotations reflect real-world understanding and use. Additionally, research has shown that “community” annotation—where non-expert annotators contribute to labeling—can be as reliable as expert annotation, expanding the pool of contributors without compromising quality. For example, researchers compared the results of teams of community annotators to accurately annotate discharge summaries from i2b2 compared to expert annotators.[61] Results showed that community annotators produced high-quality labels, nearly matching expert agreement, suggesting that crowdsourcing can reliably support complex medical annotation tasks.

3.5 Resources and Tools to Support AI Training Dataset Development and AI Model Development

We identified several resources and tools that researchers, AI developers, health care organizations, and industry can use to address data preprocessing. We also identified resources to support AI development, including data quality validation.

Data Preprocessing. There are several tools available to support de-identification (Exhibit 4). These include free open-source solutions as well as commercial options.

Exhibit 4. Example De-identification Tools and Tool Accessibility

Example De-identification Tools	Tool Accessibility
National Library of Medicine Scrubber: A clinical text de-identification tool designed to produce HIPAA-compliant de-identified health information.[62]	Available for download: **https://lhncbc.nlm.nih.gov/scrubber/**
Philter [Protected Health Information filter]: A customizable open-source tool for removing PHI from free-text clinical notes using rule-based and statistical NLP approaches.[63,64]	Available for download: **https://github.com/BCHSI/philter-ucsf**
PhysioNet de-identification: An open-source algorithm for identifying and removing HIPAA PHI and an extended set of PHI from clinical documents using pattern recognition tables and data dictionaries.[65]	Available for download: **https://www.PhysioNet.org/physiotools/deid/**
Privacy Solutions (Datavant): A range of commercially available services and solutions that meet HIPAA de-identification standards, including sharing unstructured text data using redaction, obfuscation, and Expert Determination.[66]	Services and solutions available for purchase.
Privacy Analytics (IQVIA): A variety of commercially available services and solutions for risk-based de-identification and anonymization that are HIPAA compliant.[67]	Services and solutions available for purchase.
John Snow LABS: A commercially available tool to de-identify unstructured clinical text.[68]	Services and solutions available for purchase.

Existing de-identification tools face limitations, including generalizability across health care contexts (e.g., setting), time, and expense.[69] Rule-based de-identification tools rely on regular expressions and dictionaries for pattern matching, while machine-learning approaches offer better generalization but tend to be slower. Hybrid models combining rules and machine learning have shown strong performance in removing personally identifiable information. As a result, researchers using these tools need to understand the tradeoffs between performance, speed, and cost. For example, open-source tools can be more easily customized offering opportunities for broader generalizability.

Additionally, while the HIPAA Privacy Rule specifies the personal identifiers that must be de-identified, the complexity and nuance of clinical free text introduce ambiguous cases that can be difficult for automated de-identification tools to recognize.[70] Acknowledging this challenge, and revisiting annotation guidelines, can help ensure de-identification tool performance. Knowing which identifiers are removed and why enables researchers to effectively compare de-identification tools.

There are also several tools designed to support a range of NLP tasks (e.g., NER, key word extraction) that can help interpret clinical text effectively (Exhibit 5).

Exhibit 5. Example NLP Tool and Tool Accessibility

Example NLP Tools	Tool Accessibility
ModernBERT [Bidirectional Encoder Representations from Transformers]: The most recent enhancement in BERT, **a language representation model that uses AI to understand clinical free text in context.**[71,72]	Available on GitHub: https://github.com/AnswerDotAI/ModernBERT
YAKE [Yet Another Keyword Extractor]: An unsupervised method for recognizing important keywords from a document.[73]	Available on GitHub: https://github.com/LIAAD/yake
KP-Miner: An unsupervised keyword extraction algorithm.[74]	Available on GitHub: https://github.com/INESCTEC/kep

There are also multiple efforts to enhance the AI data infrastructure available to facilitate clinical text analytics for a wide range of use cases, including comparative clinical effectiveness research (CER), clinical care, and public health surveillance, at scale. For example, the current NLP infrastructure available to researchers within the University of Florida Clinical and Translational Science Institute and the OneFlorida+ Clinical Data Research Network within PCORnet® can be leveraged for CER.[75] This NLP infrastructure includes resources such as a large clinical domain language model, GatorTron, that uses over 90 billion words of text,[76] clinical concept extraction, medical relation extraction, semantic textual similarity, natural language inference, and medical question answering, as well as tools to support computable phenotyping and disease screening. New York University (NYU) developed an LLM, NYUTron, trained on a 4.1-billion-word corpus of unstructured clinical notes. NYU researchers demonstrated that this LLM can be effectively and efficiently deployed to support a range of predictive analytics tasks across several clinical areas (e.g., 30-day all-cause readmission prediction).[77]

Additionally, researchers at Loyola University have developed two new clinical NLP analytic systems, Clinical NLP Analytic Engine and Inference Engine, which make it easier for researchers to analyze and interpret unstructured clinical text in real-time.[78] These enhanced approaches to NLP simplify how data are analyzed and how inferences are made, improving scalability and reducing the need to rely on specialized expertise that some health care organizations may lack. EHR developers are also building AI solutions directly into their products (e.g., the Epic database Cosmos). Finally, OS-PCORTF projects developed open-source NLP tools that analyze unstructured clinical data such as adverse event reports and cancer pathology reports.[79,80] These tools have been integrated into a nationwide cancer reporting system and used to create an annotated dataset for adverse event surveillance.[81]

AI Model Development. In exploring approaches to developing gold standard AI training datasets, we also identified complementary data infrastructure that supports reproducible AI models. For example, shared AI development environments can accelerate development cycles by facilitating data sharing and collaboration and promoting standardized protocols (e.g., grounded truth, data, and models). MEDS (the Medical Event Data Standard) is an open-source data standard for structured, longitudinal medical record data that supports reproducible machine learning and AI research.[82] Another approach to enhance AI model development are "model cards".[83] Model cards are concise, structured summaries that document an AI model's data sources, intended uses, training data, performance, and limitations. Encouraging the use of model cards can help

promote discoverability, transparency, and reproducibility, particularly given the limitations of existing publicly available training datasets and the reliance on proprietary or local data for AI model development.[84]

AI Model Data Quality Validation. In addition to training datasets used to build AI models, benchmark datasets are an important tool for evaluating AI performance. Several longitudinal benchmark datasets that include de-identified EHR clinical notes include INSPECT, MedAlign.85 These datasets address some of the limitations (e.g., missing longitudinal health care context) of existing training datasets like MIMIC. Similar benchmark datasets are emerging for LLMs, such as MedAgentBench.86 AI model developers can also use available training datasets (e.g., MIMIC) to evaluate algorithm robustness to sparsity and irregularity.87

4. Discussion

The development of gold standard AI training datasets—those that engage clinical domain experts, apply validated criteria, and incorporate transparent methods and feedback loops—is essential for training effective AI models.[39,88] These gold standard AI training datasets directly influence model performance by ensuring both quality and volume of training data.

We are at a pivotal inflection point where the AI landscape is evolving so rapidly that traditional methods for generating training datasets are being transformed. These advances are delivering efficiencies and, importantly, enabling AI to extract meaningful clinical insights from rich, unstructured data such as clinical notes. Researchers are quickly adopting innovative approaches, accelerating the development of high-quality AI training datasets and expanding their availability for research and clinical applications.

However, challenges remain to AI training dataset development, including data de-identification, representativeness, annotation efficiency, and the need for metrics to guide the use of synthetic data. Addressing these issues requires research that balances algorithmic generalizability with inherent trade-offs in AI model development and optimization (e.g., performance and computational cost).

Data Preprocessing. One of the biggest barriers to making AI training datasets widely available, especially at the scale needed for robust and generalizable AI model development, is the risk of re-identification. Future research can help mitigate the concern of re-identification by establishing industry-wide metrics for assessing the risk of re-identification. Researchers use metrics such as F1-scores, a single score that balances precision and recall to evaluate the performance of machine learning algorithms used to classify data. Research highlights the need for additional metrics that also assess the trade-off between de-identification and data utility.[29] Broad adoption of such standards could increase confidence in sharing data for AI development. To complement the proposed industry de-identification metrics, researchers could focus on developing benchmark datasets to assess privacy and re-identification risk in the context of creating synthetic clinical notes that can be shared for research.[23]

Another area of future AI training dataset development is the need for larger datasets that include a wider range of symptom presentations, patient profiles, and settings. This work should also focus on developing AI training datasets that incorporate multimodal data.[28] One way to increase the volume of AI training datasets that represent a range of different data types is to promote data sharing that leverages new developments in the areas of pseudonymized embedding, an AI technique that masks PHI by representing unique identifiers as numerical tokens. Broader uptake of this approach has the potential to encourage the dissemination and sharing of proprietary corpora while maintaining high levels of data privacy.[29]

Data Annotation: Advances in NLP can reduce researchers' reliance on manually annotated training corpora.[41] However, there's a need to balance the efficiency gains (e.g., needing fewer manually labeled examples) associated with using generative AI with the accuracy and precision of domain expertise.[24] To address this need, future research efforts can improve the accuracy of NLP tasks used for annotation. For example, a study evaluating the performance of LLMs for NER showed that LLMs can significantly enhance discontinuous name entity identification, yet suggested the need for further exploration of relaxed boundary annotations for improved mapping to standard ontologies.[37] This research could help reduce the need to develop custom annotation schemes, which is a time and labor-intensive activity. Evaluating the impact of single versus multiple rounds of expert annotation could further refine dataset quality and model performance. These studies could extend to include the benefit of one versus multiple rounds of expert annotation.[38]

Synthetic Dataset Creation. Given the challenges with de-identification and annotation, synthetic data have emerged as a valuable alternative for generating the volume and variety of training data. To promote continued growth in the use of gold standard synthetic AI training datasets, research is needed to establish quality measures —such as completeness, representativeness, and data heterogeneity—for synthetic dataset creation.[49,89] These metrics can help AI developers assess fitness for specific AI tasks like clinical phenotyping.

Limitations. There are three limitations to this literature review. First, most of the work on the use of NLP to analyze EHR free text data, as well as AI model development and performant evaluation, uses proprietary data. We acknowledge the substantial commercial investments in AI model development. However, while the scan includes studies conducted by individual health care organizations, lessons learned from the progress made in AI model development using institutional EHR data and home-grown solutions may be underrepresented in the literature. Second, the literature discussing best practices for AI training dataset development is limited. To address this limitation, our scan focused on AI model development. Finally, this is a rapidly evolving field, and the findings presented are based on a subset of studies that discuss AI training dataset development.

5. Conclusion

High-quality AI training datasets are foundational to developing accurate and reliable AI models for use in clinical decision-making. However, most existing datasets are proprietary and narrowly scoped, limiting reproducibility and generalizability. This report highlights the current landscape of real-world and synthetic datasets, identifies promising approaches to creating AI training datasets that have the potential to improve the AI models trained on them, and describes other resources and tools for more robust AI model development. Further, research focused on addressing challenges related to transparency, representativeness, and scalability can help improve the quality of the datasets that are used for AI development.

References

1. Office of the Assistant Secretary for Planning and Evaluation. Office of the Secretary's Patient-Centered Outcomes Research Trust Fund (OS-PCORTF): Strengthening Data Infrastructure for Patient-Centered Outcomes Research (PCOR). U.S. Department of Health and Human Services.
2. Office of the Assistant Secretary for Planning and Evaluation. Office of the Secretary Patient-Centered Outcomes Research Trust Fund (OS-PCORTF). https://aspe.hhs.gov/collaborations-committees-advisory-groups/os-pcortf
3. Alowais SA, Alghamdi SS, Alsuhebany N, et al. Revolutionizing healthcare: the role of artificial intelligence in clinical practice. BMC Med Educ. 2023;23(1):689. Published 2023 Sep 22. doi:10.1186/s12909-023-04698-z
4. Raghavan P, Chen JL, Fosler-Lussier E, Lai AM. How essential are unstructured clinical narratives and information fusion to clinical trial recruitment?. AMIA Jt Summits Transl Sci Proc. 2014;2014:218-223. Published 2014 Apr 7.
5 . The White House. Restoring the Gold Standard of Science. Published May 2025. Accessed October 28, 2025. https://www.whitehouse.gov/presidential-actions/2025/05/restoring-gold-standard-science/
6. The White House. White House unveils America's AI action plan. The White House. Published July 23, 2025. Accessed August 18, 2025. https://www.whitehouse.gov/articles/2025/07/white-house-unveils-americas-ai-action-plan/
6. Polevikov S. Advancing AI in healthcare: A comprehensive review of best practices. *Clin Chim Acta*. 2023;548:117519. doi:10.1016/j.cca.2023.117519
8. Kueper JK, Liaw W, Lizotte DJ, Tsuei SH-T. Why Is Primary Care Different? Considerations for Machine Learning Development with Electronic Medical Record Data. *NEJM AI*. 2025;2(5). doi:10.1056/AIp2401122
9. Mak S, Thomas A. Steps for Conducting a Scoping Review. J Grad Med Educ. 2022;14(5):565-567. doi:10.4300/JGME-D-22-00621.1
10. Office of the National Coordinator for Health Information Technology (ONC). Training Data for Machine Learning (ML) to Enhance Patient-Centered Outcomes Research (PCOR) Data Infrastructure: Final Report. September 2021. Accessed November 18, 2025. https://aspe.hhs.gov/sites/default/files/documents/1df150d4035554a7a1b23b0f37597e48/ONC-Training-Data-for-ML-PCOR-Final-Report.pdf
11. Ferguson S, Tille PM. Public Datasets: A Foundation to Artificial Intelligence in Health Care. Clin Lab Sci. 2024;37(2):56-63. doi:10.29074/ascls.2024003254
12. MIT Laboratory for Computational Physiology. MIMIC: Medical Information Mart for Intensive Care. Accessed November 18, 2025. https://mimic.mit.edu/
13. University of Colorado Boulder. TemporalWiki: Main Page. Accessed November 18, 2025. https://clear.colorado.edu/TemporalWiki/index.php/Main_Page
14. SyntheticMass. Synthea™: Synthetic Patient Population Simulator. Accessed November 18, 2025. https://synthetichealth.github.io/synthea/#home
15. Du J, Zeng D, Li Z, Liu J, Lv M, Chen L, Zhang D, Ji S. An interpretable outcome prediction model based on electronic health records and hierarchical attention. Int J Intell Syst. 2021 Oct 11;37(6):3460–79. doi: 10.1002/int.22697
16. Office of the National Coordinator for Health Information Technology. Synthetic Health Data Generation to Accelerate Patient-Centered Outcomes Research (PCOR): Final Report. Washington, DC: U.S. Department of Health and Human Services; March 2022.

17. Chen A, Chen DO. Simulation of a machine learning enabled learning health system for risk prediction using synthetic patient data. *Sci Rep*. 2022;12(1):17917. Published 2022 Oct 26. doi:10.1038/s41598-022-23011-4

18. Tao C, Filannino M, Uzuner Ö. FABLE: A Semi-Supervised Prescription Information Extraction System. *AMIA Annu Symp Proc*. 2018;2018:1534-1543. Published 2018 Dec 5.

19. Fraile Navarro D, Ijaz K, Rezazadegan D, et al. Clinical named entity recognition and relation extraction using natural language processing of medical free text: A systematic review. *Int J Med Inform*. 2023;177:105122. doi:10.1016/j.ijmedinf.2023.105122

20. Au Yeung J, Shek A, Searle T, et al. Natural language processing data services for healthcare providers. *BMC Medical Informatics and Decision Making*. 2024;24:356. doi:10.1186/s12911-024-02713-x

21. Sivarajkumar S, Tam TYC, Mohammad HA, et al. Extraction of sleep information from clinical notes of Alzheimer's disease patients using natural language processing. *J Am Med Inform Assoc*. 2024;31(10):2217-2227. doi:10.1093/jamia/ocae177

22. Hsu E, Malagaris I, Kuo YF, Sultana R, Roberts K. Deep learning-based NLP data pipeline for EHR-scanned document information extraction. *JAMIA Open*. 2022;5(2):ooac045. Published 2022 Jun 11. doi:10.1093/jamiaopen/ooac045

23. Chuang YS, Sarkar AR, Hsu YC, Mohammed N, Jiang X. Robust privacy amidst innovation with large language models through a critical assessment of the risks. *J Am Med Inform Assoc*. 2025;32(5):885-892. doi:10.1093/jamia/ocaf037

24. Hu Y, Chen Q, Du J, et al. Improving large language models for clinical named entity recognition via prompt engineering. *J Am Med Inform Assoc*. 2024;31(9):1812-1820. doi:10.1093/jamia/ocad259

25. Litake O, Park BH, Tully JL, Gabriel RA. Constructing synthetic datasets with generative artificial intelligence to train large language models to classify acute renal failure from clinical notes. *J Am Med Inform Assoc*. 2024;31(6):1404-1410. doi:10.1093/jamia/ocae081

26. U.S. Department of Health & Human Services. Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the HIPAA Privacy Rule. Accessed November 18, 2025. https://www.hhs.gov/hipaa/for-professionals/special-topics/de-identification/index.html

27. Burckhardt P, Padman R. deidentify. *AMIA Annu Symp Proc*. 2018;2017:485-494. Published 2018 Apr 16.

28. Wu G, Zheng L, Xie H, et al. Large Language Model Empowered Privacy-Protected Framework for PHI Annotation in Clinical Notes. *Stud Health Technol Inform*. 2025;329:876-880. doi:10.3233/SHTI250965

29. Kovačević A, Bašaragin B, Milošević N, Nenadić G. De-identification of clinical free text using natural language processing: A systematic review of current approaches. *Artif Intell Med*. 2024;151:102845. doi:10.1016/j.artmed.2024.102845

30. Dernoncourt F, Lee JY, Uzuner O, Szolovits P. De-identification of patient notes with recurrent neural networks. *J Am Med Inform Assoc*. 2017;24(3):596-606. doi:10.1093/jamia/ocw156

31. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. *arXiv*. Published February 22, 2018. Accessed November 18, 2025. https://arxiv.org/abs/1802.08232

32. Baumel T, Manoel A, Jones D, et al. Controllable synthetic clinical note generation with privacy guarantees. Published online September 12, 2024. Accessed December 16, 2025. 10.48550/arXiv.2409.07809

33. Goyal M, Mahmoud QH. A Systematic Review of Synthetic Data Generation Techniques Using Generative AI. *Electronics*. 2024; 13(17):3509. https://doi.org/10.3390/electronics13173509

34. Yan J, Huang H, Yang K, Xu H, Li Y. Synthetic data for enhanced privacy: A VAE-GAN approach against membership inference attacks. *Knowl Based Syst*. 2025;309:112899. doi:10.1016/j.knosys.2024.112899

35. Hsieh C, Moreira C, Blanco Nobre I, Costa Sousa S, Ouyang C, Brereton M, Jorge J, Nascimento JC. DALL-M: Context-Aware Clinical Data Augmentation with LLMs. arXiv [Preprint]. Published July 11, 2024. Updated March 15, 2025. doi:10.48550/arXiv.2407.08227.

36. Doumanas D, Bouchouras G, Soularidis A, Kotis K, Vouros G. From Human- to LLM-Centered Collaborative Ontology Engineering. Applied Ontology. 2025;19(4):334-367. doi:10.1177/15705838241305067

37. Alhassan A, Schlegel V, Aloud M, Batista-Navarro R, Nenadic G. Discontinuous named entities in clinical text: A systematic literature review. *J Biomed Inform*. 2025;162:104783. doi:10.1016/j.jbi.2025.104783

38. Sun W, Rumshisky A, Uzuner O. Annotating temporal information in clinical narratives. *J Biomed Inform*. 2013;46 Suppl(0):S5-S12. doi:10.1016/j.jbi.2013.07.004

39. Zhu Y, Song T, Zhang Z, Yin M, Yu P. A Five-Step Workflow to Manually Annotate Unstructured Data into Training Dataset for Natural Language Processing. *Stud Health Technol Inform*. 2024;310:109-113. doi:10.3233/SHTI230937

40. Fries JA, Steinberg E, Khattar S, et al. Ontology-driven weak supervision for clinical entity classification in electronic health records. *Nat Commun*. 2021;12(1):2017. Published 2021 Apr 1. doi:10.1038/s41467-021-22328-4

41. Zeinali N, Albashayreh A, Fan W, White SG. Symptom-BERT: Enhancing Cancer Symptom Detection in EHR Clinical Notes. *J Pain Symptom Manage*. 2024;68(2):190-198.e1. doi:10.1016/j.jpainsymman.2024.05.015

42. Lorge I, Joyce DW, Taylor N, Nevado-Holgado A, Cipriani A, Kormilitzin A. Detecting the clinical features of difficult-to-treat depression using synthetic data from large language models. Comput Biol Med. 2025;194:110246. doi:10.1016/j.compbiomed.2025.110246

43. Figueira A, Vaz B. Survey on Synthetic Data Generation, Evaluation Methods and GANs. *Mathematics*. 2022; 10(15):2733. https://doi.org/10.3390/math10152733

44. Gabriel RA, Litake O, Simpson S, Burton BN, Waterman RS, Macias AA. On the development and validation of large language model-based classifiers for identifying social determinants of health. *Proc Natl Acad Sci U S A*. 2024;121(39):e2320716121. doi:10.1073/pnas.2320716121

45. Rao H, Liu W, Wang H, Huang I, He Z, Huang X. A scoping review of synthetic data generation for biomedical research and applications. arXiv. Published June 19, 2025. doi:10.48550/arXiv.2506.16594

46. Foraker R, Morrow JD, Johnson JA, Wilcox AB, Forster AJ, Payne PRO. Understanding synthetic data: artificial datasets for real-world evidence. *BMJ Evid Based Med.* Published online July 2, 2025. doi:10.1136/bmjebm-2024-113617

47. Marchesi R, Micheletti N, I-Hsien Kuo N, Barbieri S, Jurman G, Osmani V. Generative AI mitigates representation bias and improves model fairness through synthetic health data. *PLoS Comput Biol*. 2025;21(5):e1013080. doi:10.1371/journal.pcbi.1013080

48. Engelmann J, Lessmann S. Conditional Wasserstein GAN-based oversampling of tabular data for imbalanced learning. *Expert Syst Appl*. 2021;174:114582. doi:10.1016/j.eswa.2021.114582.

49. Chen J, Chun D, Patel M, Chiang E, James J. The validity of synthetic clinical data: a validation study of a leading synthetic data generator (Synthea) using clinical quality measures. *BMC Med Inform Decis Mak*. 2019;19(1):44. Published 2019 Mar 14. doi:10.1186/s12911-019-0793-0

50. Meeker D, Kallem C, Heras Y, Garcia S, Thompson C. Case report: evaluation of an open-source synthetic data platform for simulation studies. *JAMIA Open*. 2022;5(3):ooac067. Published 2022 Aug 8. doi:10.1093/jamiaopen/ooac067

51. Qu G, Chen Q, Wei W, Lin Z, Chen X, Huang K. Mobile Edge Intelligence for Large Language Models: A Contemporary Survey. *arXiv*. Published July 2024. arXiv:2407.18921.

52. Dettmers T, Artetxe M, Shazeer N, et al. QLoRA: Efficient Finetuning of Quantized LLMs. *arXiv*. Published July 2024. doi:10.48550/arXiv.2407.18921.

53. Bellamy RKE, Dey K, Hind M, Hoffman SC, Houde S, Kannan K, et al. AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM J Res Dev*. 2019;63(4–5):4:1–4:15. doi:10.1147/JRD.2019.2942287.

54. Fairlearn. Fairlearn: Assessing and improving fairness in AI systems. Available at: https://fairlearn.org/. Accessed December 16, 2025.

55. Ethical XAI. Ethical Explainable AI. Available at: https://ethicalxai.com/index.html. Accessed December 16, 2025.

56. Pradhan S, Elhadad N, South BR, et al. Evaluating the state of the art in disorder recognition and normalization of the clinical narrative. *J Am Med Inform Assoc*. 2015;22(1):143-154. doi:10.1136/amiajnl-2013-002544

57. Li J, Zhou Y, Jiang X, et al. Are synthetic clinical notes useful for real natural language processing tasks: A case study on clinical entity recognition. J Am Med Inform Assoc. 2021;28(10):2193-2201. doi:10.1093/jamia/ocab112

58. Libbi CA, Trienes J, Trieschnigg D, Seifert C. Generating Synthetic Training Data for Supervised De-Identification of Electronic Health Records. Future Internet. 2021; 13(5):136. https://doi.org/10.3390/fi13050136

59. Styler WF 4th, Bethard S, Finan S, et al. Temporal Annotation in the Clinical Domain. Trans Assoc Comput Linguist. 2014;2:143-154.

60. Agnikula Kshatriya BS, Sagheb E, Wi CI, et al. Identification of asthma control factor in clinical notes using a hybrid deep learning model. *BMC Med Inform Decis Mak*. 2021;21(Suppl 7):272. Published 2021 Nov 9. doi:10.1186/s12911-021-01633-4

61. Uzuner Ö, Solti I, Xia F, Cadag E. Community annotation experiment for ground truth generation for the i2b2 medication challenge. *Journal of the American Medical Informatics Association*. 2010;17(5):519-523. doi:10.1136/jamia.2010.004200

62. National Library of Medicine. Scrubber: A De-identification Tool. Accessed October 28, 2025. https://lhncbc.nlm.nih.gov/scrubber/

63. Norgeot B, Muenzen K, Peterson TA, et al. Protected Health Information filter (Philter): accurately and securely de-identifying free-text clinical notes. *NPJ Digit Med*. 2020;3:57. Published 2020 Apr 14. doi:10.1038/s41746-020-0258-y

64. Radhakrishnan L, Schenk G, Muenzen K, et al. A certified de-identification system for all clinical text documents for information extraction at scale. *JAMIA Open*. 2023;6(3):ooad045. Published 2023 Jul 4. doi:10.1093/jamiaopen/ooad045

65. Neamatullah I, Douglass MM, Lehman LW, et al. Automated de-identification of free-text medical records. BMC Med Inform Decis Mak. 2008;8:32. Published 2008 Jul 24. doi:10.1186/1472-6947-8-32

66. Datavant. Connect: Privacy-Preserving Record Linkage. Accessed October 28, 2025. https://www.datavant.com/products/connect-privacy

67. Privacy Analytics. Accessed November 13, 2025. https://privacy-analytics.com/

68. John Snow LABS. Medical Data De-identification. Accessed November 20, 2025. https://www.johnsnowlabs.com/deidentification/

69. Sundrelingam V, Parimoo S, Pogacar F, et al. pyDeid: an improved, fast, flexible, and generalizable rule-based approach for deidentification of free-text medical records. *JAMIA Open*. 2025;8(1):ooae152. Published 2025 Jan 22. doi:10.1093/jamiaopen/ooae152

70. Kayaalp M, Browne AC, Sagan P, McGee T, McDonald CJ. Challenges and Insights in Using HIPAA Privacy Rule for Clinical Text Annotation. AMIA Annu Symp Proc. 2015;2015:707-716. Published 2015 Nov 5.

71. ModernBERT. GitHub. Accessed Novemebr 20, 2025. https://github.com/AnswerDotAI/ModernBERT

72. Warner B, Chaffin A, Clavié B, et al. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. *arXiv.* Published December 19, 2024. Accessed November 20, 2025. https://doi.org/10.48550/arXiv.2412.13663

73. LIAAD. YAKE: Yet Another Keyword Extractor. Accessed November 14, 2025. GitHub. https://github.com/LIAAD/yake

74. El-Beltagy SR, Rafea A. KP-Miner: A keyphrase extraction system for English and Arabic documents. Inf Syst. 2009;34(1):132-144. doi:10.1016/j.is.2008.05.002

75. University of Florida Clinical and Translational Science Institute. *Natural Language Processing (NLP) Core*. Accessed October 28, 2025. https://www.ctsi.ufl.edu/research/laboratory-services/nlp-core/

76. Yang X, Chen A, PourNejatian N, et al. A large language model for electronic health records. *NPJ Digit Med.* 2022;5(1):194. doi:10.1038/s41746-022-00742-2.

77. Jiang LY, Liu XC, Nejatian NP, et al. Health system-scale language models are all-purpose prediction engines. Nature. 2023;619(7969):357-362. doi:10.1038/s41586-023-06160-y

78. Loyola University of Chicago. Clinical NLP Analytics Engine/Inference Engine. Research Computing Services. Accessed November 20, 2025. https://www.luc.edu/its/rcs/cnaecnienlp/?

79. U.S. Department of Food and Drug Administration. VAERS-Annotations. GitHub. Accessed November 19. 2025. https://github.com/FDA/VAERS-Annotations

80. Centers for Disease Control and Prevention. NLP Workbench. GitHub. Accessed November 19, 2025. https://github.com/CDCgov/NLPWorkbench

81. Office of the Assistant Secretary for Planning and Evaluation. Development of a Natural Language Processing (NLP) Web Service for Public Health Use. Accessed November 19, 2025. https://aspe.hhs.gov/development-natural-language-processing-nlp-web-service-public-health-use

82. Medical Event Data Standard. *Medical Event Data Standard*. Accessed October 28, 2025. https://medical-event-data-standard.github.io/

83. Hugging Face. Model cards. Hugging Face Hub documentation. Published 2024. Accessed January 8, 2026. https://huggingface.co/docs/hub/en/model-cards

84. Gilbert S, Adler R, Holoyad T, Weicken E. Could transparent model cards with layered accessible information drive trust and safety in health AI? NPJ Digit Med. 2025 Feb 25;8(1):124. doi: 10.1038/s41746-025-01482-9.

85. Fries JA, Wornow M, Steinberg E, et al. Advancing Responsible Healthcare AI with Longitudinal EHR Datasets. Stanford Institute for Human-Centered Artificial Intelligence. Published February 10, 2025. Accessed October 28, 2025. https://hai.stanford.edu/news/advancing-responsible-healthcare-ai-longitudinal-ehr-datasets

86. Jiang Y, Black KC, Geng G, et al. MedAgentBench: A Virtual EHR Environment to Benchmark Medical LLM Agents. *NEJM AI*. 2025;2(9). doi:10.1056/AIdbp2500144

87. Khaled A, Sabir M, Qureshi R, Caruso CM, Guarrasi V, Xiang S, Zhou SK. Leveraging MIMIC Datasets for Better Digital Health: A Review on Open Problems, Progress Highlights, and Future Promises. arXiv. Published June 15, 2025. Accessed October 13, 2025. doi:10.48550/arXiv.2506.12808

88. JASON. Artificial Intelligence for Health and Health Care. 2017. Accessed November 18, 2025. https://www.healthit.gov/sites/default/files/jsr-17-task-002_aiforhealthandhealthcare12122017.pdf

89. Walonoski J, Kramer M, Nichols J, et al. Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. *Journal of the American Medical Informatics Association*. 2018;25(3):230-238. doi:10.1093/jamia/ocx079

Appendix A. Methods

We conducted a literature review to identify strategies, key considerations, best practices, and data infrastructure gaps in developing training datasets for AI model development. The literature review included PubMed searches for peer-reviewed articles (**Exhibit A1**) related to training datasets for NLP in EHRs.

Exhibit A1. Literature Search Terms

Search Term Category	Terms	# of Results
NLP EHR Data	(("Natural Language Processing"[MeSH Major Topic] OR "Machine Learning"[MeSH Major Topic] OR "Machine Learning Algorithms"[MeSH Major Topic]) AND ("electronic health records"[MeSH Major Topic] OR "unstructured data"[All Fields] OR "clinical notes"[All Fields] OR "free-text"[All Fields]) AND ("2015/09/11 00:00":"3000/01/01 05:00"[Date - Publication] AND "systematic review"[Filter])) AND (y_10[Filter])	44
EHR Training Data	((electronic health records[MeSH] AND ("unstructured data" OR "clinical notes" or "free-text")) AND ("training datasets" OR "training data")) Filters: from 2015 - 2025	38
Synthetic EHR Data	"Synthetic"[tiab] AND ("electronic health records"[Majr] or "health record"[tiab]) AND "natural language processing"[mesh]	21

We also reviewed materials from the following OS-PCORTF projects relevant to AI training and synthetic dataset development (**Exhibit A2**).

Exhibit A2. Relevant OS-PCORTF Projects

REMIS	Project Title
HP-19-016	A Synthetic Health Data Generation Engine to Accelerate Patient-Centered Outcomes Research (CDC & FDA)
HP-16-003	Development of a Natural Language Processing (NLP) Web Service for Public Health Use (ASTP/ONC)
HP-18-006	Enhancing Identification of Opioid-Involved Health Outcomes Using Linked Hospital Care and Mortality Data
HP-19-006	Identifying Co-Occurring Disorders among Opioid Users Using Linked Hospital Care and Mortality Data
HP-19-017	Training Data for Machine Learning to Enhance Patient-Centered Outcomes Research (PCOR) Data Infrastructure (NIH & ASTP/ONC)
HP-23-030	Utilizing Natural Language Processing and Machine Learning to Enhance the Identification of Stimulant and Opioid-Involved Health Outcomes in the National Hospital Care Survey (CDC)

Two team members conducted a title and abstract screening of 114 identified peer-reviewed articles and 44 pieces of grey literature. Articles were included or excluded based on the criteria below (**Exhibit A3**). A total of 31 peer-reviewed articles and 34 pieces of grey literature were included after full-text abstraction.

Exhibit A3. Literature Search Inclusion and Exclusion Criteria

Category	Inclusion Criteria	Exclusion Criteria
Publication Year	• Published in the last 10 years	• Published before 2015
Language	• English	• Non-English
Focus	• Study used text-based EHR clinical data • Dataset is publicly accessible	• Study does not use clinical free-text EHR data • Study used other clinical data types (e.g., imaging) • Study does not use clinical EHR data • Study focuses exclusively on the use of NLP for phenotyping, clinical event identification, prediction (without discussion of a training or synthetic dataset) • Study focuses exclusively on evaluating the performance of NLP

For the full-text abstraction, information was extracted into a spreadsheet, which provides details of the resource dataset, steps of dataset creation, and identified gaps and opportunities in the data infrastructure for training dataset development. The specific fields included in the abstraction form are shown in **Exhibit A4**.

Exhibit A4. Fields Included in Full Article Abstraction Form

Field Type/Category	Specific Fields
Resource Metadata	• Article Title • Author • Publication Year • DOI
Dataset Characteristics	• Dataset Title • Data Source • Dataset Owner • Patient Population • Access Details • Clinical/Research Use Case
NLP Training Dataset Creation	• Creation Methodology • Software Used • Best Practices for Dataset Creation • Challenges Encountered in Dataset Creation
NLP Synthetic Dataset Creation	• Creation Methodology • Software Used • Validation Criteria/Approach • Best Practices for Synthetic Dataset Creation • Challenges Encountered in Synthetic Dataset Creation

Field Type/Category	Specific Fields
Data De-identification	• Approach to Data De-Identification • Best Practices in Data De-Identification • Challenges in Data De-Identification
Data Annotation	• Annotation Techniques • Benefits of Annotation Technique • Limitations of Annotation Technique
Data Infrastructure Gaps	• Areas for Future Research • Notes

Appendix B. Publicly Available Training Datasets Identified

Sr No.	Dataset	Description
1	Acute-renal-failure-from-clinical-notes	This dataset is a collection of synthetic, labeled clinical notes created to support the detection of acute renal failure. Developed as part of a study focused on training large language models, the dataset includes annotations that enable supervised learning for identifying cases of acute renal failure from clinical documentation.
2	CACER (Clinical Concept Annotations for Cancer Events and Relations)	CACER is a corpus with fine-grained annotations for over 48,000 medical problems and drug events and 10,000 drug-problem and problem-problem relations designed to support advanced NLP tasks such as event extraction, temporal reasoning, and relation classification.
3	CheXpert Plus	CheXpert Plus is a large-scale chest radiograph dataset that contains over 223,000 images and 187,000 associated radiology reports from 64,725 patients at Stanford Hospital. It includes structured metadata, pathology labels, and RadGraph annotations, supporting multimodal research in clinical AI.
4	CLEF eHealth	CLEF eHealth is a series of shared tasks and evaluation labs held from 2013 to 2021, aimed at advancing clinical text processing. It leverages diverse datasets from U.S. and international clinical documents, including free-text from medical records and biomedical literature, to support research in information extraction, retrieval, and patient-centered technologies.
5	CLERK (Care Lifecycle Events and Records Knowledgebase)	CLERK Protege is a commercial solution consisting of encounter-level clinical data linked with claims data for tens of millions of encounters, built specifically for training AI models in health care administration tasks (e.g., billing and coding).
6	EHR-DS-QA: A Synthetic QA Dataset Derived from Medical Discharge Summaries for Enhanced Medical Information Retrieval Systems	EHR-DS-QA is a synthetic question-answering dataset derived from 21,466 medical discharge summaries in the MIMIC-IV-Note database. It contains 156,599 Q&A pairs generated using a custom pipeline built on the 13-billion parameter LLaMA-2 model. The dataset was developed to support the training of healthcare-focused large language models for enhanced clinical information retrieval and decision support. A subset of the Q&A pairs was verified by a physician for accuracy.
7	Healthcare Dataset	This synthetic healthcare dataset is a simulated dataset designed to reflect real-world healthcare scenarios, enabling users to practice data analysis, machine learning, and data science techniques. It includes structured patient information and clinical variables, supporting experimentation and skill development in healthcare analytics.
8	Healthverity Notes	HealthVerity Notes is a modular, de-identified dataset comprising over 2.5 billion unstructured clinical notes sourced from electronic health records (EHRs) across ambulatory and multi-specialty practices nationwide.
9	INSPECT EHR (Integrating Numerous Sources for Prognostic Evaluation of Clinical Timelines)	INSPECT is a longitudinal electronic health record (EHR) benchmark dataset that contains de-identified clinical data from 19,390 patients diagnosed with pulmonary embolism (PE). It includes 23,248 computed tomography pulmonary angiography (CTPA) studies, each paired with radiology impressions, and provides ground truth labels for multiple clinical outcomes. The dataset supports research in prognostic modeling, temporal event prediction, and multimodal learning by integrating structured EHR data with imaging-derived insights.

Sr No.	Dataset	Description
10	MDClone ADAMS Platform	MDClone ADAMS Platform is a commercially available self-service healthcare data analytics environment designed to support exploration, research, and innovation across clinical and operational domains. It enables secure access to structured and unstructured data from hospital EHR systems, preserving patient privacy through synthetic data generation.
11	MediSyn Inc	MediSyn is a high-fidelity, syntehtic EHR dataset generated using advanced machine learning techniques. MediSyn can produce large, longitudinal, datasets customized to specific conditions or patient cohorts, enabling targeted research and model development. The dataset supports integration with the Observational Medical Outcomes Partnership common data model and Fast Healthcare Interoperability Resources.
12	MedNLI	MedNLI is a publicly available, clinically grounded dataset designed for natural language inference (NLI) tasks in the medical domain. It consists of de-identified sentence pairs derived from patient discharge summaries written by clinicians. The dataset is tailored for evaluating and training models on clinical reasoning and textual entailment, making it a valuable resource for developing machine learning systems that understand and interpret clinical language.
13	MIMIC (Medical Information Mart for Intensive Care)	MIMIC is an acute and critical care database that contains detailed clinical data on over 360,000 patients admitted to Beth Israel Deaconess Medical Center (BIDMC) between 2001 and 2022 and the Northwestern ICU database, a critical care dataset collected from a network of hospitals located in Chicago and harmonized with the MIMIC data structure. The Northwestern ICU database includes de-identified data from Northwestern Memorial HealthCare from 2020-2022. The latest version, MIMIC-IV, consists of six linkable modules (Hospital, ICU, Diagonstic Electrocardiogram, ED, Chest X-Ray, Note). The Hospital and ICU modules contain data from both BIDMC and Northwestern Medicine Healthcare hospitals. Earlier versions include MIMIC-III and MIMIC-II, which cover earlier timeframes.
14	MMIT	The MMIT dataset includes over 2.5 billion de-identified clinical notes sourced from EMRs across 600 hospitals and 40 health systems, transformed into longitudinal, analytics-ready data. These notes encompass physician dictations, specialty consultations, pathology reports, and patient-reported symptoms.
15	MTSamples	The MTSamples dataset is a collection of thousands of sample medical transcriptions across a wide range of specialties and clinical domains, such as cardiology, neurology, surgery, psychiatry, and emergency medicine. The dataset is commonly used in clinical NLP research for tasks like text classification, named entity recognition, and specialty prediction. It is available under a CC0 Public Domain license, making it freely usable for research and educational purposes.
16	n2c2 NLP Research Data Sets	The National NLP Clinical Challenges (n2c2) is a series of shared tasks and evaluation labs held from 2006 onward, designed to advance natural language processing for clinical narratives. Originating from the i2b2 initiative, n2c2 provides de-identified clinical notes from U.S.-based medical records, supporting research in information extraction, concept recognition, and relation classification across nine topics: de-identification and smoking, obesity, medications, relations, coreference, temporal relations, de-identification and hear disease, clinical trial cohort selection, and adverse drug events and medication extraction.

Sr No.	Dataset	Description
17	NYU Notes	NYU Notes is a clinical text dataset composed of notes authored by licensed medical professionals across multiple NYU Langone hospitals from 2011 to 2020.
18	PsyTar (Psychiatric Treatment Adverse Reactions)	PsyTAR is a corpus developed to support research in psychiatric pharmacovigilance and clinical NLP. It contains patient-generated drug reviews from askapatient.com, focusing on SSRIs and SNRIs. The dataset includes annotated sentences and entities related to adverse drug reactions, withdrawal symptoms, drug indications, and effectiveness. This corpus consists of three main components: sentence classification, entity identification, and terminology association.
19	SemEval (Semantic Evaluation)	The SemEval datasets are a collection of benchmark datasets used in a long-running series of shared tasks focused on the evaluation of computational semantic analysis systems. Each year's SemEval workshop introduces new tasks and datasets, often reflecting emerging challenges in natural language processing. These datasets are manually annotated and designed to test systems on nuanced aspects of meaning, context, and reasoning in language.
20	SHAIP	Shaip EHR is a comprehensive electronic health records dataset containing over 5.1 million de-identified medical records and 257,000+ hours of physician dictation audio across 31 specialties. Curated from diverse healthcare settings across the United States, the dataset includes structured and unstructured clinical data such as patient histories, diagnoses, prescriptions, treatment plans, radiology images, and laboratory results.
21	ShARe (Shared Annotated Resources) Project	The ShARe corpus is a dataset compiled for clinical NLP tasks using medical transcriptions. It consists of deidentified clinical free-text notes from the MIMIC II database. The corpus supports research in disorder recognition, normalization, and other clinical information extraction tasks. It has been used in shared evaluations such as CLEF eHealth and SemEval, and includes gold standard annotations designed to advance interoperability and standardization in clinical NLP.
22	Subsalt	Subsalt is a synthetic data platform designed to generate HIPAA-compliant replicas of healthcare datasets for research, prototyping, and clinical AI development.
23	Syntegra Synthetic EHR Data Structured Healthcare Electronic Health Record Data	Syntegra is a commercially available synthetic dataset developed to replicate the statistical distribution of clinical data from U.S.-based hospital EHR systems. It was created to support the development and evaluation of machine learning models across a wide range of clinical tasks.
24	Synth Clinical Notes	Synth Clinical Notes is a synthetic dataset of clinical notes developed from GPT-4o-mini containing information related to housing eviction. It was developed to evaluate the performance of the SynthEHR-Eviction pipeline in identifying social determinants of health (SDOH) related to eviction within unstructured clinical text.
25	Synthea	Synthea™ is a synthetic patient generator that models the medical history of synthetic patients across all aspects of healthcare, including demographics, encounters, procedures, medications, and outcomes. Synthea supports export in multiple formats (e.g., FHIR, CSV, OMOP CDM) and enables population-level simulations based on geographic and epidemiological parameters.

Sr No.	Dataset	Description
26	THYME (Temporal Histories of Your Medical Event)	THYME is a corpus developed to support temporal information extraction from clinical text. It includes annotated pathology, radiology, and clinical notes focused on brain and colon cancer, using a schema based on TimeML. The dataset enables research in event detection, temporal relation extraction, and timeline construction from clinical narratives. The THYME Project includes three corpora (2015, 2016, 2017).
27	Veradigm	Veradigm EHR is a large-scale electronic health record (EHR) dataset containing detailed clinical data on over 150 million unique patients across the United States, collected from diverse healthcare settings between Q3 2020 and Q2 2025. The dataset integrates structured and unstructured data—including diagnoses, medications, procedures, lab results, and over 5 billion clinical notes—sourced from multiple EHR systems.