

Comprehensive Analysis of Open-Access Longitudinal Electronic Health Record Resources for Oncologic Research

Executive Summary

The intersection of oncology, longitudinal patient care, and unstructured clinical data represents one of the most fertile yet operationally challenging frontiers in biomedical informatics. For researchers and data scientists, the ability to access open, anonymized, and longitudinal Electronic Health Record (EHR) datasets is fundamental to developing robust Artificial Intelligence (AI) and Natural Language Processing (NLP) models capable of parsing the complex trajectory of cancer treatment. This report provides an exhaustive analysis of the current landscape of such datasets, with a specific focus on resources that include unstructured clinical notes—the repository of critical oncologic details such as tumor staging, progression, and treatment rationale often missing from structured data.

The analysis reveals that while the demand for high-quality, longitudinal oncology data is immense, the supply is constrained by significant privacy and interoperability hurdles. The landscape is currently dominated by a single, monolithic resource—**MIMIC-IV (Medical Information Mart for Intensive Care)**—which serves as the *de facto* standard for open-access critical care research, including significant cohorts of cancer patients. Unlike other datasets that offer only snapshots or structured claims data, MIMIC-IV provides linked, longitudinal visits with rich unstructured narratives (discharge summaries and radiology reports) for hundreds of thousands of patients.

Beyond MIMIC-IV, the ecosystem is supported by specialized NLP challenge datasets (n2c2/i2b2) which offer high-quality annotations for longitudinal tasks, albeit on smaller cohorts. Furthermore, a burgeoning field of synthetic data generation, led by tools like **Synthea**, offers a privacy-preserving alternative for modeling ideal oncologic pathways (breast, lung, colorectal) and generating synthetic clinical notes. Innovative approaches such as the **DFCI-cancer-outcomes-ehr** "Teacher-Student" framework demonstrate new paradigms where private institutional data trains models to label public datasets, effectively transferring clinical knowledge without transferring Protected Health Information (PHI).

This report details the structural, semantic, and longitudinal characteristics of these datasets. It provides researchers with the technical specifications needed to extract cancer cohorts, the legal frameworks governing access (PhysioNet Credentialed Access), and the methodological insights required to leverage unstructured text for predictive oncology. The

findings underscore that while "perfect" longitudinal oncology data remains rare in the public domain due to the complexity of outpatient follow-up, the integration of critical care datasets (MIMIC-IV) with advanced synthetic modeling and NLP transfer learning offers a viable pathway for substantial research progress.

1. The Landscape of Oncologic Data Availability

1.1 The "Unstructured" Imperative in Oncology

Oncology is inherently narrative. While structured data fields in EHRs capture billing codes (ICD-10, CPT) and medication orders, they fail to capture the nuance of cancer care. The trajectory of a cancer patient is defined by a series of complex decisions and biological evolutions that are recorded primarily in free-text documents.

Phenotypic Depth in Narratives:

- **Pathology Reports:** These contain the definitive diagnosis, including histological subtypes (e.g., "adenocarcinoma with lepidic pattern"), tumor grade, and increasingly, molecular biomarkers (e.g., EGFR mutation status, PD-L1 expression). Structured fields rarely capture this level of granularity.¹
- **Radiology Reports:** These are the primary source for determining disease status. Terms like "interval increase in size of hepatic metastasis" or "new osseous lesion" define disease progression (PD) or response (PR/CR) according to RECIST criteria. This information is almost never structured.³
- **Clinical Notes (Admission/Discharge/Progress):** These synthesize the patient's status, documenting toxicity (e.g., "grade 3 neuropathy"), performance status (ECOG/Karnofsky), and the rationale for treatment changes (e.g., "discontinuing oxaliplatin due to neurotoxicity").

Therefore, for a dataset to be truly valuable for oncology research, it must be **multimodal** (structured + text) and **longitudinal** (linking visits over time to track disease evolution). The user's requirement for "unstructured clinical notes" is not merely a preference but a prerequisite for meaningful oncologic inquiry.⁵

1.2 The "Longitudinal" Challenge

Cancer is a chronic disease characterized by phases of care: diagnosis, curative intent therapy, surveillance, recurrence, palliative therapy, and end-of-life care. Capturing this full arc in an open-access dataset is exceptionally difficult due to data fragmentation.

- **Fragmentation:** A patient might be diagnosed at a community hospital, have surgery at an academic medical center, receive chemotherapy at a local infusion center, and enter hospice at home.

- **Data Silos:** Integrating these disparate records into a single longitudinal file requires interoperability that even healthcare systems struggle to achieve internally, let alone in public datasets.⁷

The Definition of "Longitudinal" in Open Data:

In the context of this report, "longitudinal" refers to datasets that meet three criteria:

1. **Multiple Time Points:** The data must capture more than a single snapshot of care.
2. **Patient Linkage:** A persistent pseudo-identifier must link these time points across different encounters (e.g., subject_id linking admission A and admission B).
3. **Temporal Integrity:** The relative timing of events must be preserved, even if absolute dates are shifted for privacy.⁸

1.3 Privacy Barriers and the "Open-Access" Paradox

The user's query highlights a critical tension: the need for open-access data versus the need for granular, longitudinal data. Longitudinal narratives are notoriously difficult to de-identify. A sequence of dates, rare cancer subtypes, and specific treatment complications can re-identify a patient even if names are removed.

- **HIPAA Safe Harbor:** To share data publicly, 18 identifiers must be removed, including all dates more granular than the year. For oncology, where "progression-free survival" is measured in days or months, this loss of temporal precision is a major hurdle.
- **The "Mosaic Effect":** Re-identification risk increases with the amount of data available. A longitudinal record with detailed notes is a high-risk asset.

Despite these barriers, the open-science community, primarily through platforms like **PhysioNet**, has established robust frameworks (credentialed access, rigorous de-identification) to make such data available. This report focuses on these credentialed resources that balance openness with privacy.⁵

2. The Primary Resource: MIMIC-IV (Medical Information Mart for Intensive Care)

MIMIC-IV stands as the unparalleled giant in the landscape of open-access health data. While its name implies a focus on intensive care, its scope has expanded in version IV to include hospital-wide data, making it a viable, albeit biased, source for oncology research. It is currently the only open dataset that satisfies all the user's criteria: longitudinal, real-world, anonymized, and rich in unstructured text.¹⁰

2.1 Architecture and Modules

MIMIC-IV is sourced from the Beth Israel Deaconess Medical Center (BIDMC) in Boston, MA. It

is organized into modules to separate data by source and granularity.

2.1.1 The hosp Module

This module contains data derived from the hospital-wide EHR (distinct from the ICU system). It is the backbone for longitudinal analysis.

- **patients Table:** The core registry. It provides the subject_id, anchor_year (a shifted year to preserve privacy), and anchor_age. Crucially, it links to state death records, providing **out-of-hospital mortality** labels up to one year post-discharge. This is vital for oncology survival analysis.¹⁰
- **admissions Table:** Tracks every hospitalization. A single subject_id can be linked to multiple hadm_id entries, allowing researchers to construct a sequence of visits (e.g., diagnostic workup → surgical admission → admission for febrile neutropenia).
- **diagnoses_icd Table:** Contains the coded diagnoses for each admission. This is the primary entry point for cohort selection (identifying cancer patients via ICD-9/10 codes).

2.1.2 The note Module (MIMIC-IV-Note)

This is the most critical component for the user's request. It contains the unstructured narratives linked to the hosp and icu modules.¹¹

- **discharge Table:** Contains ~331,794 discharge summaries. These are comprehensive narratives summarizing a patient's entire hospital stay. For cancer patients, these summaries often include a "History of Present Illness" that recaps the entire cancer journey up to that point, mitigating the lack of outpatient data.
- **radiology Table:** Contains ~2.3 million radiology reports. These reports are the ground truth for tumor staging and response assessment. They detail the findings of CT scans, MRIs, and X-rays, using language that is essential for NLP tasks like "detecting metastasis" or "measuring tumor burden."

2.1.3 The icu Module

While less relevant for general oncology, this module provides high-resolution physiological data (vitals, drips) for cancer patients who become critically ill.

2.2 The Oncology Cohort Within MIMIC-IV

MIMIC-IV is not an oncology registry; it is a critical care and hospital admission dataset. However, due to the sheer volume of patients (~364,000), it contains a massive "hidden" oncology cohort. Cancer patients frequently interface with the hospital for complications, surgeries, or terminal care.¹²

2.2.1 Cohort Identification Strategy

Researchers typically identify cancer patients by querying the diagnoses_icd table for codes

starting with:

- **ICD-9:** 140-239 (Neoplasms)
- **ICD-10:** C00-D49 (Neoplasms)

Volume Estimates:

- **Lung Cancer:** Studies have identified cohorts of **1,755 to 2,500+** unique lung cancer patients with longitudinal data.¹³
- **Sepsis in Cancer:** A study identified **3,796** cancer patients admitted with sepsis.¹⁵
- **General Oncology:** Broad queries typically yield over **10,000+** unique patients with some form of malignancy documented in their structured data and notes.

2.2.2 Characteristics of the Cohort

The MIMIC oncology population is skewed towards:

- **Acuity:** Patients are admitted, often to the ICU, implying advanced disease or severe complications (e.g., sepsis, respiratory failure).
- **Mortality:** In-hospital mortality is higher than the general cancer population (e.g., ~21% for lung cancer admissions).¹⁴
- **Comorbidities:** The data is rich in comorbidity information (heart failure, renal failure), allowing for complex risk stratification modeling.

2.3 MIMIC-IV-Ext-22MCTS: The Temporal Layer

One of the most significant recent developments for longitudinal research is the release of **MIMIC-IV-Ext-22MCTS**.⁹ This "Extended" dataset addresses the difficulty of parsing raw text for timelines.

The Problem: A discharge summary might say, "Patient started chemo 3 days ago, developed fever yesterday." A machine learning model needs this as a structured time-series.

The Solution: This dataset uses Large Language Models (LLMs) to extract **22 million clinical events** from the MIMIC-IV discharge summaries.

- **Structure:** It provides a table with columns for Event (text description), Time (relative hours since admission), and Time_bin (discrete intervals).
- **Utility:** This effectively converts the unstructured narrative into a **structured longitudinal event stream**. Researchers can use this to model the precise trajectory of a cancer patient's hospitalization (e.g., Admission → Imaging → Chemo → Reaction → Antibiotics) without building their own extraction pipeline from scratch.

Table 1: Comparison of Core MIMIC-IV vs. Extended Modules

Feature	MIMIC-IV Core	MIMIC-IV-Note	MIMIC-IV-Ext-22 MCTS
Data Type	Structured (Codes, Vitals)	Unstructured (Text)	Structured Time-Series (Text + Time)
Source	Hospital EHR	Transcription Systems	Derived from Notes via LLM
Primary Use	Mortality Prediction, Epi	Phenotyping, NLP	Temporal Modeling, Trajectory Analysis
Longitudinality	Linked Admissions	Linked Notes	Event-level Granularity
Access	PhysioNet Credentialed	PhysioNet Credentialed	PhysioNet Credentialed

2.4 Limitations for Oncology

While MIMIC-IV is the best available resource, researchers must be aware of its limitations:

- Missing Outpatient Data:** MIMIC captures inpatient stays. The routine "every 3 weeks" outpatient chemotherapy cycles are often missing unless summarized in a history note.
- Date Shifting:** To protect privacy, dates are shifted to the future (2100s). While intervals (days between visits) are preserved, the absolute seasonality (e.g., flu season) and correlation with external events (e.g., FDA approval of a new drug in 2015) are lost.
- Censoring:** Patients may receive care at other hospitals (e.g., Dana-Farber) that is not captured in the BIDMC dataset, leading to incomplete treatment histories.

3. Specialized and Derived Resources

Beyond the monolithic MIMIC-IV, there are specialized datasets and frameworks designed to address specific challenges in oncology NLP, such as the need for high-quality labels ("Does this note indicate progression?") which raw EHR data lacks.

3.1 The DFCI "Teacher-Student" Framework

This resource, **DFCI-cancer-outcomes-ehr**, represents a paradigm shift in how sensitive

oncology data is shared.¹⁷ It addresses the problem that expert-labeled oncology notes (e.g., "This patient has RECIST Progression") are almost never public.

3.1.1 The Mechanism

1. **The Teacher:** Researchers at Dana-Farber Cancer Institute (DFCI) trained "Teacher" models (based on hierarchical transformers and Clinical-Longformer) on their *private*, identified data. These models learned to extract outcomes like **Disease Progression**, **Response**, and **Metastasis** from radiology reports and oncologist notes.
2. **The Distillation:** These Teacher models were then applied to the *public* MIMIC-IV dataset to generate labels.
3. **The Student:** "Student" models were trained on the MIMIC-IV text using the Teacher-generated labels.

3.1.2 The Resource

The "dataset" released on PhysioNet is not just rows of data, but the **Student Models** themselves.

- **Utility:** A researcher can download these models and apply them to the MIMIC-IV-Note dataset. The result is a MIMIC-IV dataset *enriched* with high-quality, expert-derived oncology labels.
- **Performance:** The student models achieved high discrimination (AUC > 0.90) for detecting progression and metastasis, effectively transferring the clinical expertise from a specialized cancer center to a general public dataset.

3.2 n2c2 and i2b2 Clinical Challenges

The **National NLP Clinical Challenges (n2c2)**, formerly i2b2, provide smaller but rigorously annotated datasets. Unlike MIMIC, where labels must be inferred or extracted, n2c2 datasets come with "Gold Standard" human annotations.¹⁸

3.2.1 2014 Longitudinal Clinical Narratives

This track is the most relevant for the user's request for "longitudinal" data.²⁰

- **Content:** ~1,300 documents for ~300 patients.
- **Structure:** Each patient has 2 to 5 notes spread over time.
- **Focus:** While the primary domain was diabetes and heart disease, the dataset is invaluable for testing **temporal reasoning** in NLP. The challenge required systems to track risk factors over time (e.g., "Did the patient smoke *before* the diagnosis?"). This logic is directly transferable to oncology (e.g., "Did the patient receive radiation *before* surgery?").
- **Access:** Available via the **DBMI Data Portal** (Harvard Medical School) after signing a DUA.

3.2.2 2018 Cohort Selection

This track focused on identifying patients who met specific criteria for clinical trials.²⁰

- **Relevance:** Clinical trial matching is a major use case for oncology AI. This dataset provides a benchmark for systems that must read longitudinal notes to determine eligibility (e.g., "History of malignancy", "Current steroid use").

3.3 Limitations of Other "Open" Datasets

The analysis of research snippets identified several other datasets that often appear in searches but fail to meet the "longitudinal" or "patient-linked" criteria.

- **MTSamples (Kaggle/MTSamples.com):** This is a collection of transcribed medical reports.²³
 - Pros: Completely open, no DUA required. Contains "Oncology" specific reports.
 - Cons: **No Patient ID.** Each row is an isolated document. There is no way to link a pathology report to a subsequent surgery note. It is useful for training language models on medical jargon but useless for longitudinal patient modeling.
- **HuggingFace Snippets:** Various "Synthetic Clinical Note" datasets exist on HuggingFace (e.g., starmpcc/Asclepius-Synthetic-Clinical-Notes).²⁵
 - Pros: Easy access.
 - Cons: Often generated by LLMs (GPT-4) without grounding in a real patient trajectory. They lack the noise, temporal inconsistency, and complexity of real EHRs.

4. Synthetic Data Ecosystems

Given the scarcity of real-world longitudinal oncology text, **synthetic data** has emerged not just as a stopgap, but as a strategic asset. These datasets are computer-generated to statistically resemble real patients, avoiding privacy risks entirely.

4.1 Synthea and mCode Modules

Synthea is an open-source patient generator that models the "life history" of synthetic patients.²⁷ It uses a state-machine approach where a "patient" moves between states (Health → Disease → Treatment) based on transition probabilities derived from epidemiology.

4.1.1 Oncology Specificity

Synthea includes specific modules for **Breast Cancer**, **Colorectal Cancer**, and **Lung Cancer** based on the **mCode** (minimal Common Oncology Data Elements) standard.²⁸

- **Mechanism:** The software simulates the entire cancer journey: screening (mammography), diagnosis (biopsy), staging (TNM), treatment (chemo/rads/surgery),

and survival.

- **Longitudinality:** It generates data from "cradle to grave." A synthetic patient will have a birth date, primary care visits, cancer diagnosis, chemotherapy cycles, and death outcomes.
- **Unstructured Text:** Synthea can generate simple clinical notes (Subjective, Objective, Assessment, Plan) derived from templates. While they lack the linguistic messiness of real dictations, they contain **perfect ground truth** for entities. If the note says "Stage III," the structured data is Stage III.

4.1.2 The "Coherent Data Set"

A specific release of Synthea data, the **Coherent Data Set**, includes linked FHIR resources, DICOM images, and clinical notes.³⁰ This is designed to test interoperability and multimodal AI models.

4.2 Generative AI Approaches

Emerging research utilizes Large Language Models (LLMs) to generate synthetic notes that are more realistic than Synthea's templates.

- **MedGemma / GPT-4:** Researchers prompt these models with a structured patient profile (e.g., "60yo male, lung cancer, progressed on chemo") to generate a realistic-sounding discharge summary.³¹
- **Utility:** These datasets are often released as small benchmarks on GitHub/HuggingFace to test NLP extraction performance without exposing real PHI.

5. Technical Implementation Strategies

For a researcher possessing these datasets, the challenge shifts from *acquisition* to *operationalization*. Mining longitudinal oncology insights from MIMIC-IV or Synthea requires specific technical strategies.

5.1 Phenotyping and Cohort Selection

Identifying the cancer cohort in MIMIC-IV is the first step. Relying solely on ICD codes is often insufficient due to coding errors or the presence of "history of" codes (e.g., a patient with a past cured cancer).

Recommended SQL Strategy:

1. **Broad Sweep:** Query hosp.diagnoses_icd for all 140-239 (ICD-9) and C00-D49 (ICD-10) codes.
2. **Filter by Sequence:** Use the seq_num field to prioritize primary diagnoses (seq_num = 1) over secondary ones.
3. **NLP Validation:** Join with note.discharge and use regex to search the "History of Present

- "Illness" section for keywords like "metastatic", "stage IV", "chemotherapy", or "radiation".
4. **Pathology Proxies:** Since MIMIC lacks a structured path table, extracting "GLEASON SCORE" (Prostate) or "ER+/PR+" (Breast) from the text helps confirm active disease.³³

5.2 Extracting Outcomes: RECIST and Survival

The **Response Evaluation Criteria in Solid Tumors (RECIST)** is the standard for assessing cancer treatment (Complete Response, Partial Response, Stable Disease, Progressive Disease).

- **Data Source:** Radiology Reports in MIMIC-IV.
- **Method:** Apply the **DFCI Student Models** (see Section 3.1) to classify each radiology report.
- **Longitudinal Construction:** By stringing these labels together over time, researchers can construct a **Time-to-Progression (TTP)** curve.
 - *Visit 1 (CT Chest): "Target lesion 2.0 cm" → Baseline.*
 - *Visit 2 (CT Chest): "Target lesion 1.5 cm" → Partial Response.*
 - *Visit 3 (CT Chest): "Target lesion 3.0 cm" → Progressive Disease.*

5.3 Handling "Date Shifting" in Survival Analysis

MIMIC-IV's date shifting preserves intervals but destroys absolute time.

- **Calculations:** Survival must be calculated as `date_of_death - date_of_diagnosis`. Both dates are shifted by the same amount for a given patient, so the *difference* is accurate.
- **Seasonality Bias:** Researchers cannot study seasonal effects (e.g., "Are cancer patients more likely to get pneumonia in January?") because "January" in the dataset might actually be "July" in reality.
- **External Correlation:** You cannot correlate MIMIC data with external real-world events (e.g., "Did mortality drop after the approval of Keytruda in 2014?") because the years are shifted to 2100+.

6. Future Directions and Recommendations

The landscape of open-access longitudinal oncology data is evolving rapidly. The static "download a CSV" model is being replaced by "bring your code to the data" models and synthetic generation.

6.1 The Rise of Federated Learning

Projects like **Rhino Health** and **ODHSI/OMOP** are pushing for federated learning, where models travel to hospitals, train on private data, and return only the weights. This avoids the

need to de-identify and share raw longitudinal notes.³⁴

6.2 Recommendations for the User

For the specific user query—"open-access, longitudinal, anonymized, English, oncology, with notes"—the following roadmap is recommended:

1. **Primary Asset: MIMIC-IV** (PhysioNet). It is the only resource with the necessary volume and longitudinal linkage.
 - o *Action:* Complete CITI training, sign the DUA, and download the hosp and note modules.
2. **Text Enrichment: MIMIC-IV-Note.**
 - o *Action:* Use this for the raw narratives.
3. **Temporal Structure: MIMIC-IV-Ext-22MCTS.**
 - o *Action:* Use this derived dataset to jump-start temporal modeling without needing to build a custom event extractor.
4. **Labeling: DFCI Student Models.**
 - o *Action:* Use these pre-trained models to "annotate" the MIMIC radiology reports with RECIST criteria.
5. **Synthetic Supplement: Synthea (mCode).**
 - o *Action:* Use this if the research requires "idealized" outpatient trajectories (e.g., screening to diagnosis) that are missing from MIMIC's inpatient-heavy data.

Table 2: Strategic Resource Selection Matrix

Research Goal	Recommended Dataset	Reason
Survival Analysis	MIMIC-IV (hosp + patients)	Links to death records; calculating survival time is possible.
NLP Model Training (NER)	n2c2 2014 / Synthea	High-quality annotations or perfect ground truth.
Disease Progression Modeling	MIMIC-IV-Ext-22MCTS	Pre-extracted time-series events allow for trajectory modeling.
Clinical Trial Matching	n2c2 2018	Specifically designed for cohort selection tasks.
Multimodal AI (Text +	MIMIC-IV (Radiology)	Links text reports to DICOM

Image)		images (available in MIMIC-CXR).
---------------	--	----------------------------------

By leveraging this ecosystem of primary, derived, and synthetic resources, researchers can effectively overcome the barriers of privacy and fragmentation to conduct meaningful longitudinal oncology research.

7. Ethical and Legal Frameworks

7.1 PhysioNet Credentialing

Access to MIMIC-IV and its derivatives is governed by the **PhysioNet Credentialed Health Data License**. This is a "trust-based" model that relies on researcher accountability.

- **Requirement:** Users must complete the **CITI Data or Specimens Only Research** course (ethics training). This ensures researchers understand the sensitivity of the data.
- **Agreement:** Users sign a Data Use Agreement (DUA) explicitly forbidding:
 - **Re-identification:** Any attempt to link the data to public records (e.g., obituaries, voter rolls) to find patient identities.
 - **Redistribution:** Posting the data on GitHub, Kaggle, or public drives.
- **Audit:** PhysioNet audits user activity. Violations result in bans and potential legal action.

7.2 Cloud Access and "Data Sandboxes"

There is a growing trend toward **Cloud-Native Access** (e.g., Google BigQuery, AWS Open Data).

- **Mechanism:** Instead of downloading 1TB of data to a local hard drive (high risk), researchers query the data directly in a secure cloud environment.
- **Benefit:** This enhances security (data never leaves the controlled environment) and democratizes access (researchers don't need massive local servers). MIMIC-IV is fully integrated into Google BigQuery, allowing for SQL-based analysis of the entire longitudinal record without a single download.

7.3 The Ethics of Synthetic Data

Synthetic data (Synthea) bypasses HIPAA entirely because there are no "real" patients to protect.

- **Pros:** Can be shared freely, hosted on GitHub, and used for education.
- **Cons:** It introduces **Model Bias**. If the underlying model (e.g., transition probabilities in Synthea) assumes that "all Stage IV lung cancer patients receive chemo," the data will reflect that, potentially erasing the disparities (e.g., rural patients not accessing care) seen in real-world data like MIMIC. Researchers must acknowledge this "idealized reality"

when publishing results.

8. Detailed Technical Appendix: MIMIC-IV Schema for Oncology

To assist the user in operationalizing the MIMIC-IV dataset, this section provides a conceptual map of the tables required to build a longitudinal cancer dataset.

Figure 1: Conceptual Schema for Oncology Extraction in MIMIC-IV

- **Patient Core:**
 - subject_id (Unique Patient)
 - gender
 - anchor_age (Age at a shifted date)
 - dod (Date of Death - crucial for survival analysis)
- **Hospitalization Layer** (Linked by subject_id):
 - hadm_id (Unique Admission)
 - admittime / dischtime
 - insurance / marital_status (Social determinants)
- **Clinical Data Layer** (Linked by hadm_id):
 - diagnoses_icd (Filter for 140-239 / COO-D49)
 - procedures_icd (Filter for chemotherapy/surgery codes)
 - pharmacy (Chemotherapy orders)
- **Narrative Layer** (Linked by hadm_id):
 - note.discharge (Full summary of the cancer course)
 - note.radiology (RECIST assessment source)
- **Temporal Layer** (Linked by hadm_id):
 - transfers (Movement between wards, e.g., Oncology Ward → ICU)

Example Workflow:

1. **SELECT** subject_id **FROM** diagnoses_icd **WHERE** icd_code **LIKE** 'C34%' (Lung Cancer).
2. **JOIN** with admissions to get the sequence of visits.
3. **JOIN** with note.discharge to get the text.
4. **APPLY** NLP to text to extract "Stage" and "Biomarkers".
5. **CALCULATE** Survival = dod - min(admittime).

This schema demonstrates that while MIMIC-IV is not "labeled" as an oncology dataset, it possesses all the relational structures necessary to function as one.

Works cited

1. A Question-and-Answer System to Extract Data From Free-Text Oncological Pathology Reports (CancerBERT Network): Development Study - Journal of

- Medical Internet Research, accessed on February 6, 2026,
<https://www.jmir.org/2022/3/e27210/>
- 2. Automatic Classification of Cancer Pathology Reports: A Systematic Review - PMC - NIH, accessed on February 6, 2026,
<https://pmc.ncbi.nlm.nih.gov/articles/PMC8860734/>
 - 3. An open-source framework for end-to-end analysis of electronic health record data - PMC, accessed on February 6, 2026,
<https://pmc.ncbi.nlm.nih.gov/articles/PMC11564094/>
 - 4. Empirical evaluation of artificial intelligence distillation techniques for ascertaining cancer outcomes from electronic health records - PubMed Central, accessed on February 6, 2026, <https://pmc.ncbi.nlm.nih.gov/articles/PMC12152177/>
 - 5. EpistasisLab/ClinicalDataSources: Open or Easy Access Clinical Data Sources for Biomedical Research - GitHub, accessed on February 6, 2026,
<https://github.com/EpistasisLab/ClinicalDataSources>
 - 6. OMNY Health Dataset Hits 100 Million Patient Milestone, Unlocking Unprecedented Access to Healthcare Data for Research, accessed on February 6, 2026,
<https://omnyhealth.com/omny-health-dataset-hits-100-million-patient-milestone-unlocking-unprecedented-access-to-healthcare-data-for-research/>
 - 7. Advancing Responsible Healthcare AI with Longitudinal EHR Datasets | Stanford HAI, accessed on February 6, 2026,
<https://hai.stanford.edu/news/advancing-responsible-healthcare-ai-longitudinal-ehr-datasets>
 - 8. Mapping Patient Trajectories using Longitudinal Extraction and Deep Learning in the MIMIC-III Critical Care Database* - Pacific Symposium on Biocomputing (PSB) 2026, accessed on February 6, 2026,
<https://psb.stanford.edu/psb-online/proceedings/psb18/beaulieu-jones.pdf>
 - 9. MIMIC-IV-Ext-22MCTS: A 22 Millions-Event Temporal Clinical Time-Series Dataset with Relative Timestamp - PhysioNet, accessed on February 6, 2026,
<https://physionet.org/content/mimic-iv-ext-22mcts/>
 - 10. MIMIC-IV v3.1 - PhysioNet, accessed on February 6, 2026,
<https://physionet.org/content/mimiciv/>
 - 11. MIMIC-IV-Note: Deidentified free-text clinical notes v2.2 - PhysioNet, accessed on February 6, 2026, <https://www.physionet.org/content/mimic-iv-note/2.2/>
 - 12. Characteristics and clinical subtypes of cancer patients in the intensive care unit: a retrospective observational study for two large databases, accessed on February 6, 2026, <https://pmc.ncbi.nlm.nih.gov/articles/PMC7859733/>
 - 13. Assessing in-hospital mortality risk in ICU lung cancer patients using machine learning: An analysis based on the MIMIC-IV database - Research journals - PLOS, accessed on February 6, 2026,
<https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0341259>
 - 14. Assessing in-hospital mortality risk in ICU lung cancer patients using machine learning: An analysis based on the MIMIC-IV database - PMC, accessed on February 6, 2026, <https://pmc.ncbi.nlm.nih.gov/articles/PMC12826459/>
 - 15. A nomogram for predicting hospital mortality of critical ill patients with sepsis and

- cancer: a retrospective cohort study based on MIMIC-IV and eICU-CRD | BMJ Open, accessed on February 6, 2026,
<https://bmjopen.bmj.com/content/13/9/e072112>
- 16. MIMIC-IV-Ext-22MCTS: A 22 Million-Event Temporal Clinical Time-Series Dataset for Risk Prediction - arXiv, accessed on February 6, 2026,
<https://arxiv.org/html/2505.00827v1>
 - 17. Shareable Artificial Intelligence to Extract Cancer Outcomes from Electronic Health Records for Precision Oncology Research - PhysioNet, accessed on February 6, 2026, <https://physionet.org/content/dfc1-cancer-outcomes-ehr/1.0.0/>
 - 18. n2c2 NLP Research Data Sets, accessed on February 6, 2026,
<https://portal.dbmi.hms.harvard.edu/projects/n2c2-nlp/>
 - 19. NLP Data Sets - i2b2: Informatics for Integrating Biology & the Bedside, accessed on February 6, 2026, <https://www.i2b2.org/NLP/DataSets/>
 - 20. Cohort selection for clinical trials: n2c2 2018 shared task track 1 - PMC, accessed on February 6, 2026, <https://PMC.ncbi.nlm.nih.gov/articles/PMC6798568/>
 - 21. Clinical trial cohort selection using Large Language Models on n2c2 Challenges - arXiv, accessed on February 6, 2026, <https://arxiv.org/html/2501.11114v1>
 - 22. Data Sets | National NLP Clinical Challenges (n2c2) - Harvard University, accessed on February 6, 2026, <https://n2c2.dbmi.hms.harvard.edu/data-sets>
 - 23. Medical Transcriptions - Kaggle, accessed on February 6, 2026,
<https://www.kaggle.com/datasets/tboyle10/medicaltranscriptions>
 - 24. Clinical Text Data Categorization and Feature Extraction Using Medical-Fissure Algorithm and Neg-Seq Algorithm - PMC, accessed on February 6, 2026,
<https://PMC.ncbi.nlm.nih.gov/articles/PMC8920702/>
 - 25. stamppcc/Asclepius-Synthetic-Clinical-Notes · Datasets at Hugging Face, accessed on February 6, 2026,
https://huggingface.co/datasets/stamppcc/Asclepius-Synthetic-Clinical-Notes/vie_wer
 - 26. Raymond-dev-546730/Synthetic-LungNotes-10K · Datasets at, accessed on February 6, 2026,
<https://huggingface.co/datasets/Raymond-dev-546730/Synthetic-LungNotes-10K>
 - 27. SynthNotes: A Generator Framework for High-volume, High-fidelity Synthetic Mental Health Notes - OSTI.gov, accessed on February 6, 2026,
<https://www.osti.gov/servlets/purl/1507868>
 - 28. Downloads | Synthea - Mitre, accessed on February 6, 2026,
<https://synthea.mitre.org/downloads>
 - 29. Synthea Synthetic Data Overview – FHIR® for Research Documentation, accessed on February 6, 2026,
<https://mitre.github.io/fhir-for-research/modules/synthea-overview>
 - 30. The “Coherent Data Set”: Combining Patient Data and Imaging in a Comprehensive, Synthetic Health Record - MDPI, accessed on February 6, 2026,
<https://www.mdpi.com/2079-9292/11/8/1199>
 - 31. Enhancing Cancer Symptom Detection in EHR Clinical Notes - PMC - NIH, accessed on February 6, 2026,
<https://PMC.ncbi.nlm.nih.gov/articles/PMC12433187/>

32. Medical Hallucination in Foundation Models and Their Impact on Healthcare - medRxiv, accessed on February 6, 2026,
<https://www.medrxiv.org/content/10.1101/2025.02.28.25323115v2.full.pdf>
33. A Frame-Based Nlp System For Cancer-Related Information Extraction - DigitalCommons@TMC, accessed on February 6, 2026,
https://digitalcommons.library.tmc.edu/cgi/viewcontent.cgi?article=1100&context=uthshis_docs
34. Data Analysis on MIMIC-IV about Mortality Prediction - ZHAW, accessed on February 6, 2026,
https://www.zhaw.ch/storage/engineering/institute-zentren/cai/studentische_arbeiten/Herbst_2023/1st_Master_Project_23_bogo_DataAnalysisMortalityPrediction.pdf