Enhanced Prognostic Prediction in Acute Myeloid Leukemia: Applying Graph Neural Networks to Longitudinal Electronic Health Record Data[1]

Riya Sinha, David Kuo, Ben Viggiano, Ethan Schonfeld, Matthew Schwede[2]

Department of Biomedical Data Science, Stanford University, 450 Jane Stanford Way Stanford, CA 94305, USA
Email: mschwede@stanford.edu

Acute myeloid leukemia (AML) is a life-threatening disease that affects 20,000 patients in the United States each year, and only 30% of patients live longer than five years. To determine which patients require more aggressive therapy, physicians need risk stratification methods. However, AML is highly heterogeneous, and current approaches focus on the genotype without using detailed phenotypic or longitudinal clinical information. Our group has previously applied natural language processing to sequential pathology reports to predict prognosis in AML, and this system improved on standard risk stratification. We aimed to further improve this classification system by incorporating additional data from the electronic health record. We extracted both genotype and phenotype data from the electronic health record and leveraged the longitudinal nature of those data to predict survival in AML. Specifically, we used mutations, parsed karyotypes, labs, treatment, and important descriptors from pathology reports from 868 patients and applied a graph neural network to predict overall survival at one year from the diagnosis date. This achieved an AUROC on the validation set of 0.81. This approach compared favorably to the standard risk stratification criteria using a comparable Stanford dataset (AUC 0.60). Overall, we found that using longitudinal data, additional variables beyond the conventional mutations, and modern graph neural network techniques improved the accuracy of prognosis prediction for AML.

Keywords: acute myeloid leukemia; deep learning; electronic health records; graph neural networks; heterogeneous graph transformer; weak supervision

[1] This work is supported by the Stanford Department of Biomedical Data Science T15 grant

[2] Work supported by the incredible instruction of Christine Yeh, Dennis Wall, Barbara Engelhardt, and Russ Altman. Thank you all! We had a great time in the course. 👩🏽‍💻Riya 👨🏻‍💻 David 🧙🏻‍♂️Ben 👨🏼‍🍼Ethan 👨🏼‍🦲 Matt

1. Introduction

Acute myeloid leukemia (AML) has a poor prognosis, with a five-year overall survival of 30%.[1] Recently, the array of treatments for AML has expanded substantially, with the United States Food and Drug Administration approving nine therapies between 2017 and 2020,[2] many of which target specific mutations.[3–5] This revolution in AML care has been fueled in part by analyses of large genomic datasets that have made clear the epidemiology of targetable mutations and the risk associated with each mutation.[6,7] However, even with these new treatments and greater knowledge of the disease, the prognosis of AML remains poor. Understanding which patients will benefit from each treatment is crucial, but guidelines and prognostic calculators largely ignore patient characteristics, the leukemia phenotype, and changes during a patient's course.

In clinical practice, limited information, such as the presence of individual mutations at diagnosis, is used in formal guidelines and published models to predict prognosis in AML (Table 1).[8] However, the clinical information available to hematologists and oncologists is growing rapidly and provides an opportunity to predict more accurately a patient's prognosis and response to treatment. In addition to mutations, which can predict response to selected treatments,[9–11] several other variables are routinely gathered when diagnosing, treating, and monitoring patients with AML. These include variables related to leukemia phenotype, such as the leukemia's histological characteristics, as well as patient-level variables, such as laboratory information. Although manually recording these data is laborious and error-prone, these data are often available in the electronic health record (EHR), and by more fully characterizing the leukemia and the patient, they provide additional information that could improve clinicians' predictive abilities.

Table 1. Prognostic mutations for AML from standard guidelines

Favorable	- t(8;21)(q22;q22.1)/*RUNX1*::*RUNX1T1* - inv(16)(p13.1q22) or t(16;16)(p13.1;q22)/*CBFB*::*MYH11* - Mutated *NPM1*†,§ without *FLT3*-ITD - bZIP in-frame mutated *CEBPA*
Intermediate	- Mutated *NPM1*†,§ with *FLT3*-ITD - Wild-type *NPM1* with *FLT3*-ITD (without adverse-risk genetic lesions) - t(9;11)(p21.3;q23.3)/*MLLT3*::*KMT2A* - Cytogenetic and/or molecular abnormalities not classified as favorable or adverse
Adverse	- t(6;9)(p23.3;q34.1)/*DEK*::*NUP214* - t(v;11q23.3)/*KMT2A*-rearranged - t(9;22)(q34.1;q11.2)/*BCR*::*ABL1* - t(8;16)(p11.2;p13.3)/*KAT6A*::*CREBBP* - inv(3)(q21.3q26.2) or t(3;3)(q21.3;q26.2)/*GATA2*, *MECOM(EVI1)* - t(3q26.2;v)/*MECOM(EVI1)*-rearranged - −5 or del(5q); −7; −17/abn(17p) - Complex karyotype,** monosomal karyotype - Mutated *ASXL1*, *BCOR*, *EZH2*, *RUNX1*, *SF3B1*, *SRSF2*, *STAG2*, *U2AF1*, and/or *ZRSR2* - Mutated *TP53*

In addition to using limited types of information to predict prognosis, physicians also generally rely on information from a single pretreatment time point,[8] and although gathering longitudinal data manually is similarly laborious, the EHR allows rapid extraction of longitudinal data. In other fields of medicine, longitudinal clinical data from the EHR has been used to predict outcomes, but such research is absent in AML. For example, data from the Medical Information Mart for Intensive Care (MIMIC-III) database[12] has previously been leveraged to predict mortality and other tasks in patients in the intensive care unit using the first 48 hours of data.[13] Additionally, modeling serial data related to treatment response using machine learning has been previously shown to improve prognostic and predictive computational models in other hematologic cancers.[14] Continuously updating predictions and using prior information about previous treatments and disease characteristics is especially important in AML, given that it evolves over time[15,16] and that this clonal evolution contributes to relapse.[17,18]

Thus, we set out to model prognosis in AML by leveraging variables from the EHR that are not normally incorporated into prognostic calculators for AML. We also use the longitudinal aspect of these variables for the prediction. To do this, we use several different methods to extract key variables from the electronic health record and multiple machine learning models to predict survival at one year. In our prior work, we found that the standard mutation-based prognostic scoring system resulted in an AUROC of 0.6 for survival at one year, and we found that these approaches substantially improved on the predictive abilities of standard clinical risk.

2. Methods

2.1. *Extracting data from the electronic health record*

The data from this project come from the AML Database at Stanford, a branch of the Stanford Cancer Institute Research Database. Briefly, data were collected by querying relationship databases corresponding to Stanford's electronic health record. These include Clarity, which at Stanford has also been mapped to a more concise data model, OMOP (Observational Medical Outcomes Partnership).[19,20] Both databases were used. Some variables came from structured data, such as laboratory values like the white blood cell count, hemoglobin, platelets, and neutrophil count. However, other features had to be extracted from unstructured text, such as bone marrow biopsy pathology reports.

For the purposes of this project, we identified two features we wanted to extract from the unstructured text using a mix of rule-based and weakly supervised methods. First, we defined a categorical variable named "cellularity status" to indicate whether bone marrow biopsies were specified as *hypocellular*, *normocellular*, or *hypercellular* – standard descriptors in pathology reports to indicate how much space in the bone marrow is occupied by blood cells, a measure that both varies with treatment and reflects initial disease burden. Second, we defined a binary variable named "dysplasia status," which was used to indicate if a pathology report referenced the presence of dysplasia – abnormal blood cell development which generally suggests an antecedent blood disorder. A hematologist labeled 1,000 randomly selected notes to serve as a development (n=600) and test (n=400) sets for these feature extraction methods. The open source MedSpaCy[21] was utilized to split the text from the reports into sentences and identify relevant vocabulary from controlled lists. We then iteratively designed labeling functions and created majority vote classifiers for both tasks, using the development set to characterize performance. If we were able to achieve sufficient performance for a particular task with the rule-based methods alone, the predictions of the majority vote classifier were utilized for the final feature assignments for each note. If we were unable to achieve sufficient performance for a particular task, we used the intermediate majority vote classifier to create a pseudo-labeled training set (n=15,584) - a dataset that utilizes the output of the majority vote classifier as a training label. Then, utilizing the large pseudo-labeled training set, we fine tuned a Clinical-Longformer[22] model to perform the classification task. Clinical-Longformer was selected due to its knowledge-enriched pretraining on the MIMIC-III dataset's clinical notes[12], and it's ability to receive up to 4,096 tokens as input, which enable us to pass the full text of the majority of our notes (>99%). The test set was never observed nor used to validate the label functions, majority vote classifiers, or the Clinical-Longformer end model until the final classifiers had been selected for each task to avoid data leakage.

Finally, the diagnosis date for a patient was specifically extracted from the Stanford Cancer Registry, which is a manually curated registry about patients whose cancer was either diagnosed or treated at Stanford. Retrieving accurate diagnosis dates from the medical record was challenging, and although the Cancer Registry otherwise has limited data on AML, these dates were previously found to be highly accurate in a manual review performed by a hematologist (~98% accurate to within two days).

2.2. *Building a heterogeneous graph dataset*

2.2.1. *Inclusion/exclusion criteria*

Our objective was to determine the value of longitudinal data in prognosis prediction. Consequently, our analysis only focused on patients with longitudinal data available and a recorded survival status for the pertinent time period. We analyzed data gathered within the initial six months following their diagnosis to forecast one-year survival outcomes. Therefore, any patients who passed away prior to the six-month mark, or for whom the most recent survival data was before the one-year timeframe, were not included in the study. This resulted in a dataset comprising patient information within the first six months post-diagnosis, paired with the binary outcome of either surviving or not surviving at the one-year mark.

2.2.2. *Preprocessing pipeline*

Feature-extracted data were stored in .txt and .csv files, with one file for each data type (demographics, mutations, chromosomal abnormalities, labs, chemotherapy, pathology reports). For each data type, data was loaded into a Pandas DataFrame and then preprocessed by first filtering for patients who met the inclusion and exclusion criteria and then dropping any data points collected greater than 6 months from diagnosis date, and prior to 90 days since the diagnosis date. Missing values were imputed using backwards and forwards imputation if any values were present, or mean-filling otherwise. DataFrames for each data type were then merged by patient ID to create the final preprocessed dataset for heterogeneous patient graph generation.

2.2.3. *Graph dataset construction*

In order to unify the data sources comprising the patient history, we sought a data representation that could naturally handle these multiple visit types with disjoint feature sets. Heterogeneous graphs are networks of different types of nodes and edges. Thus, each patient visit can be represented as a node in a heterogeneous graph, with a node type corresponding to its visit type. The data corresponding to the visit node can be set as the node's feature vector.

When specifying the edges in this graph, however, we are presented with an opportunity for feature engineering. Naively, the nodes in the graph can be thought of as occurring in a linear sequence of events, based on the time at which they occur. However, we can leverage clinical insights to design a more interesting graph structure. AML is thought of as evolving on the order of weeks. Thus, we specify two different edge-types. The first is a bidirectional edge between events that occur

within one week of each other. This edge represents events that occur in roughly the same time period, and helps form clusters, or chains of continuous change. The second is a unidirectional edge from the last node of each type in a cluster, to the first node in the next cluster of each type. These are intended to represent larger jumps in time which may see more drastic changes in patient state. The edge feature vectors contain a single element representing the difference in days between the linked events.

Finally, some graph-level attributes are also added, such as the age at diagnosis of the patient, the death status of the patient, and their patient ID.

2.3. *Graph neural network model architecture and training*

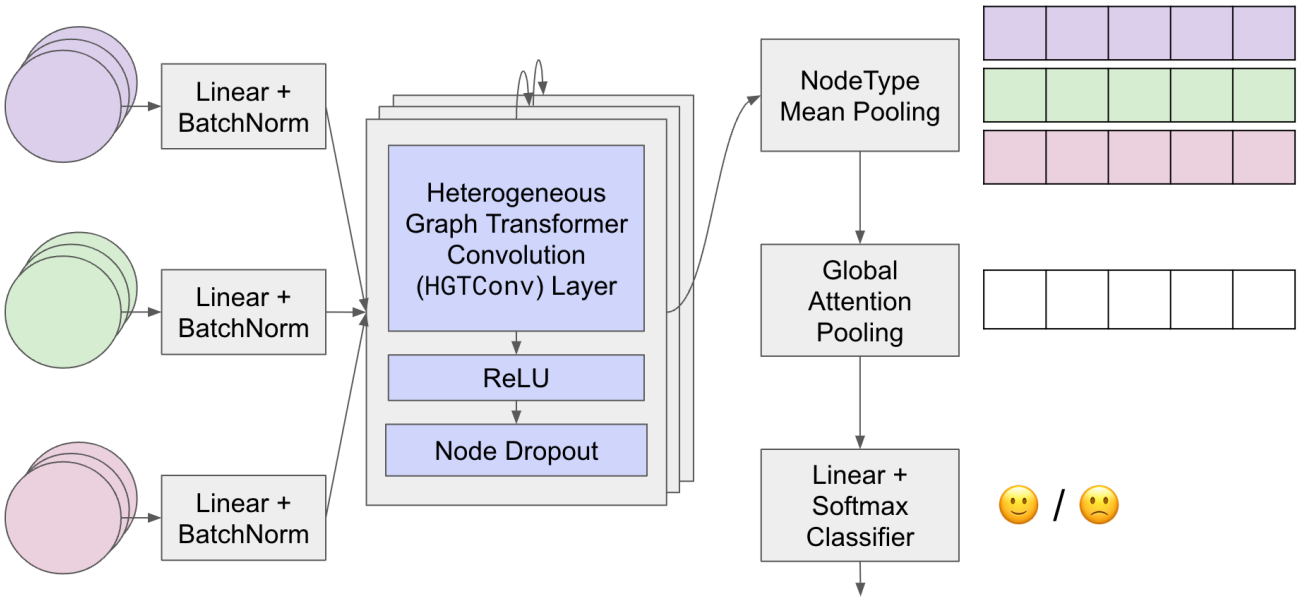

Figure 1: Graph neural network model architecture.

2.3.1. *Model architecture*

For each patient graph, we project nodes from each data type into a shared embedding space with a data type specific linear transformation followed by batch normalization. We then use a custom layer consisting of a Heterogeneous Graph Transformer Convolutional Layer followed by ReLU non-linearity and drop out with probability 0.1 to generate embeddings for each node. This custom layer is applied five times. Mean pooling is then performed on the node embeddings by node type to create an average embedding for each node type in the graph. This is followed by global attention

pooling over the node-type embeddings to produce a graph-level embedding for each patient. This graph-level embedding is then passed to a linear classification layer to predict survival at 1 year from the diagnosis date.

2.3.2. *Model training*

The model was trained end to end with binary cross entropy loss for 1000 epochs with the Adam optimizer, and weight decay of 1e-5. Hyperparameters we adjusted for were:

1. Batch size (256, 1024)
2. Learning rate (1e-3, 1e-4, 1e-5) with linear warm-up for 150 epochs
3. Hidden dimension (512, 1024, 2048)
4. Number of attention heads (4, 8, 16, 32, and 56)
5. Number of GNN layers (2-5)
6. Censorship thresholds (3 months vs. 6 months)

3. Results

3.1. *EHR feature extraction*

Structured data were extracted from the Clarity and OMOP databases corresponding to Stanford's electronic health record. Data was extracted for 2,612 patients, with all dates abstracted to be relative as "days since diagnosis" based on dates found from the Stanford Cancer Registry for each patient. Extracting structured features from these databases resulted in datasets for labs, cancer-directed therapy, chromosome abnormalities, and mutations.

Additional variables were extracted from unstructured text in pathology reports. For the "cellularity status" feature extraction task, three simple labeling functions were created. Descriptions of these label functions and their corresponding performances are included in Table A1 in the appendix. Given we were able to achieve an accuracy of 0.99 and macro F1 score of 0.99 on the development set with the majority vote classifier, we decided to utilize the majority vote classifier's predictions instead of training Clinical-Longformer for the cellularity status task. The majority vote classifier achieved an accuracy of 0.99 and macro F1 score of 0.99 on the held out test set.

For the dysplasia status task, we created five labeling functions which are described in Table A2 in the appendix. Operating on the outputs of the label functions, we created a majority vote classifier that was able to achieve an accuracy of 0.9 and F1 score of 0.82 on the development set. We then

utilized this classifier to create a pseudo-labeled training set which was used to finetune a Clinical-Longformer classification model. For the dysplasia status task, the Clinical-Longformer model was able to achieve an AUROC of 0.97, AUPRC of 0.93, an accuracy of 0.95 and F1 score of 0.88 on the held out test set, using a decision threshold of 0.5.

3.2. *Building a heterogeneous graph dataset*

From 2,612 patients for whom we had survival data, we curated a heterogenous graph dataset of 868 patients meeting our study criteria. Of these, 332 patients were deceased by 1 year of their diagnosis date and 536 were alive at 1 year of their diagnosis date.

A graph was then constructed for each of these patients saved as .pt pytorch object files. Below in Figure 2 are example visualizations of two such graphs.

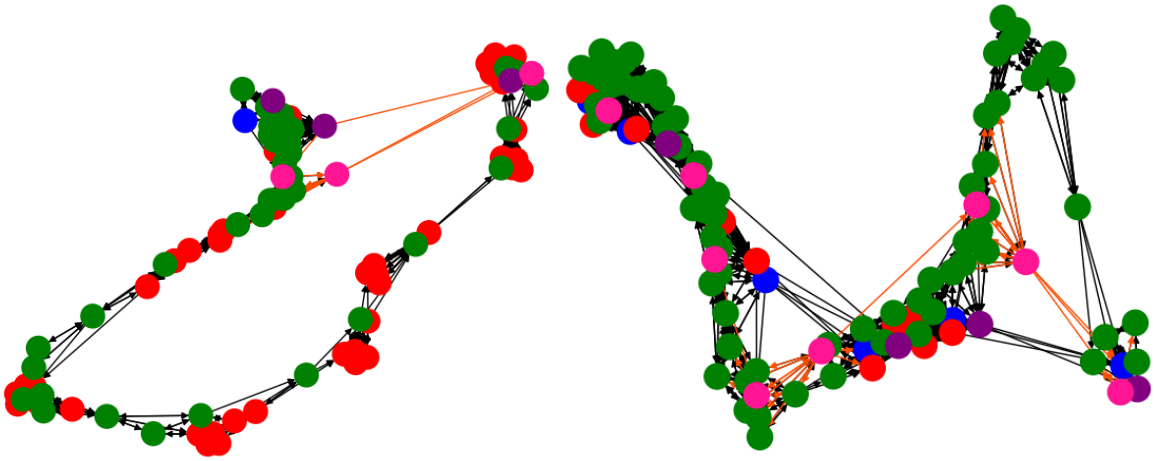

Figure 2: Visualizations of heterogeneous graphs for two patient histories. Nodes include labs (green), chemotherapy treatments (red), gene mutation testing (purple), chromosomal abnormality tests (blue), and biopsy pathology reports (pink). Edges include in-cluster (black) and out-of-cluster (orange) connections.

3.3. *Graph neural network model performance*

Our best-performing graph neural network model had 16 attention heads, 5 layers, a hidden channel dimension of 1024, was trained on 6 months of patient data, and achieved an AUROC of 81.0% and AUPRC of 72.0% in the validation set, outperforming a previously developed NLP model for predicting 1 year survival from bone marrow biopsy reports.

Table 2. Graph Neural Network Classification Performance. Accuracy, F1, Precision, and Recall were calculated using a decision threshold of 0.50.

Model	Accuracy	F1	Precision	Recall	AUROC	AUPRC
NLP	0.702	0.617	0.581	0.658	0.693	*[3]
GNN	0.733	0.623	0.760	0.528	0.810	0.721

4. Discussion

In this study, we leveraged weak supervision to develop a deep learning model for extracting relevant features from bone marrow biopsy pathology reports. Representing each patient's longitudinal data as a heterogenous graph, we then trained a graph neural network to predict survival at one year, achieving a high AUROC and AUPRC.

We demonstrated that some often-overlooked variables, such as the presence of dysplasia in the non-leukemia cells or the bone marrow cellularity, can be extracted with high accuracy from pathology reports. This is likely in part due to the consistency of the language used in pathology reports at Stanford to describe these concepts. An improvement to our system to help with generalization across other healthcare systems and reduce dependence on potentially brittle rule-based methods would be to leverage generative large language models, such as medAlpaca[23], to extract data. We did try to access the medAlpaca models from HuggingFace for this project, but unfortunately the model's tokenizer was not uploaded correctly, so we continued with our rule-based and weak supervision methods.

For the purposes of this project, we utilized per-feature imputation rules designed using domain area expertise. To ensure the validity of our imputation assumptions, we want to further investigate the processes that generated our datasets to understand how missingness manifests for various features and in what manner. It is likely that some features have missingness that is either non-random or that has correlation with other features, which could significantly impact model performance.

We furthermore found that modeling longitudinal patient data as a heterogeneous graph and training a graph neural network to predict 1 year survival was a promising direction, and one that can

[3]AUPRC was not calculated for the NLP model. As of June 1st, 2023 the publicly accessible pretrained weights of the GatorTron model were removed from HuggingFace. In the future, we plan to utilize a different model such as ClinicalLongformer.

be broadly applicable across clinical medicine. In particular, representing patient data as a graph addressed several of the challenges that our team faced when trying to represent multimodal longitudinal data. We originally tried to represent data in a concatenated tabular format to have a unified feature set, but found that this led to a large degree of sparsity for sequence models (e.g. recurrent neural networks). Similarly, applying classical models (e.g. decision trees) to longitudinal data requires summarizing the patient history in a way that inevitably loses information. By representing data as a graph, we could maintain the richness and context of the data sequence, thereby significantly enhancing model performance.

Our current graph neural network model utilizes heterogeneous graph transformers which can take advantage of the multiple node and edge types in the graph. However, a limitation of these models is that they are not able to take into account edge attributes, such as distance in time between events in our graph. Our classifier is also not using the graph-level features such as age at diagnosis at this time. Despite these limitations, our initial findings hold promise. Adding additional variables such as temporal indication of a bone marrow transplant and flow cytometry data may help improve its predictive power further. Additionally, our model performance should be compared against other time-series analysis techniques.

Given that we only have access to Stanford EHR data and labs, our analysis is limited to patients who presented for care at Stanford. To understand the generalizability of our methods, we would need to perform our analysis at other institutions, as aspects such as text feature extraction would likely be significantly impacted by various reporting writing practices. Furthermore, it is essential to perform a fairness and bias analysis of our model to understand how the predictions of various patient demographics are impacted and to ensure that our methods are capable of making accurate predictions across diverse populations.

Finally, we believe that our problem, improving prognosis prediction for acute myeloid leukemia, is one where there is a clear path for machine learning to tangibly improve the current standard of care. We envision a future in which machine learning models are seamlessly integrated into the electronic health record and automatically extract relevant clinical features and provide meaningful predictions and insights to guide clinical management.

Appendix

Table A1. Label functions for the cellularity status classification task. Coverage indicates the fraction of examples where a particular label function assigns a label (ei. does not abstain). Accuracy represents the fraction of examples that are correctly classified by each label function.

Description	Coverage	Accuracy
If a note only contains one mention of **hypercellular**, **normocellular**, or **hypocellular**: assign that class. If multiple mentioned, **abstain.**	0.735	0.99
If a note does not contain any reference to cellularity, label as **no information**. Otherwise, **abstain.**	0.24	1
If a note contains multiple cellularity status, assign whatever status if found in the Diagnosis section. If multiple are found, assign highest cellularity status. Otherwise, **abstain**.	0.025	0.99

Table A2. Label functions for the dysplasia status classification task. Coverage indicates the fraction of examples where a particular label function assigns a label (ei. does not abstain). Accuracy represents the fraction of examples that are correctly classified by each label function.

Description	Coverage	Accuracy
If **dysplasia** is mentioned more than **4** times in a note, label as **positive**. Otherwise, **abstain**.	0.135	0.907
If semi-structured table is identified, look at the next 15 tokens and determine if any are "Yes". If "Yes" is found, label as **positive**. Otherwise, label as **negative**.	0.07	0.82
If either **multi-cellular dysplasia** or **single-cellular dysplasia** is mentioned, label as **positive**. Otherwise, label as **abstain**.	0.115	0.84
If **dysplasia** is never mentioned, label as **negative**. Otherwise, **abstain**.	0.33	0.99
If **dysplasia** mentions are found, and there is at least one non-negated, non-historical, and non-contemplative reference, label as **positive**. If all **dysplasia** mentions are negated, historical, or contemplative, label as **negative**. If there are none, **abstain**.	0.27	0.93

References

1. Acute Myeloid Leukemia - Cancer Stat Facts. *SEER* https://seer.cancer.gov/statfacts/html/amyl.html.

2. Carter, J. L. *et al.* Targeting multiple signaling pathways: the new approach to acute myeloid leukemia therapy. *Signal Transduct. Target. Ther.* **5**, 288 (2020).

3. Stein, E. M. *et al.* Enasidenib in mutant IDH2 relapsed or refractory acute myeloid leukemia. *Blood* **130**, 722–731 (2017).

4. DiNardo, C. D. *et al.* Durable Remissions with Ivosidenib in IDH1-Mutated Relapsed or Refractory AML. *N. Engl. J. Med.* **378**, 2386–2398 (2018).

5. Stone, R. M. *et al.* Midostaurin plus Chemotherapy for Acute Myeloid Leukemia with a FLT3 Mutation. *N. Engl. J. Med.* **377**, 454–464 (2017).

6. Tyner, J. W. *et al.* Functional Genomic Landscape of Acute Myeloid Leukemia. *Nature* **562**, 526–531 (2018).

7. The Cancer Genome Atlas Research Network. Genomic and Epigenomic Landscapes of Adult De Novo Acute Myeloid Leukemia. *https://doi.org/10.1056/NEJMoa1301689* (2013) doi:10.1056/NEJMoa1301689.

8. Döhner, H. *et al.* Diagnosis and management of AML in adults: 2017 ELN recommendations from an international expert panel. *Blood* **129**, 424–447 (2017).

9. Stahl, M. *et al.* Clinical and molecular predictors of response and survival following venetoclax therapy in relapsed/refractory AML. *Blood Adv.* **5**, 1552–1564 (2021).

10. DiNardo, C. D. *et al.* Molecular patterns of response and treatment failure after frontline venetoclax combinations in older patients with AML. *Blood* **135**, 791–803 (2020).

11. Zhang, H. *et al.* Integrated analysis of patient samples identifies biomarkers for venetoclax efficacy and combination strategies in acute myeloid leukemia. *Nat. Cancer* **1**, 826–839 (2020).

12. Johnson, A. E. W. *et al.* MIMIC-III, a freely accessible critical care database. *Sci. Data* **3**, 160035 (2016).

13. Harutyunyan, H., Khachatrian, H., Kale, D. C., Ver Steeg, G. & Galstyan, A. Multitask learning and benchmarking with clinical time series data. *Sci. Data* **6**, 96 (2019).

14. Kurtz, D. M. *et al.* Dynamic Risk Profiling Using Serial Tumor Biomarkers for Personalized Outcome Prediction. *Cell* **178**, 699-713.e19 (2019).

15. Miles, L. A. *et al.* Single-cell mutation analysis of clonal evolution in myeloid malignancies. *Nature* 1–6 (2020) doi:10.1038/s41586-020-2864-x.

16. Jan, M. & Majeti, R. Clonal evolution of acute leukemia genomes. *Oncogene* **32**, 135–140 (2013).

17. Quek, L. *et al.* Clonal heterogeneity of acute myeloid leukemia treated with the IDH2 inhibitor enasidenib. *Nat. Med.* **24**, 1167–1177 (2018).

18. Morita, K. *et al.* Clonal evolution of acute myeloid leukemia revealed by high-throughput single-cell genomics. *Nat. Commun.* **11**, 5327 (2020).

19. OMOP Common Data Model. https://ohdsi.github.io/CommonDataModel/.

20. Belenkaya, R. *et al.* Extending the OMOP Common Data Model and Standardized Vocabularies to Support Observational Cancer Research. *JCO Clin. Cancer Inform.* **5**, 12–20 (2021).

21. Eyre, H. *et al.* Launching into clinical space with medspaCy: a new clinical text processing toolkit in Python. *AMIA. Annu. Symp. Proc.* **2021**, 438–447 (2022).

22. Li, Y., Wehbe, R. M., Ahmad, F. S., Wang, H. & Luo, Y. Clinical-Longformer and Clinical-BigBird: Transformers for long clinical sequences. Preprint at https://doi.org/10.48550/arXiv.2201.11838 (2022).

23. Han, T. *et al.* MedAlpaca -- An Open-Source Collection of Medical Conversational AI Models and Training Data. Preprint at https://doi.org/10.48550/arXiv.2304.08247 (2023).