Predicting Survival in Acute Myeloid Leukemia from Pathology Reports

Stanford CS224N Custom Project
Mentor: Elaine Sui

Name: Matthew Schwede MD
SUNet ID: mschwede
Department of Biomedical Data Science
Stanford University
mschwede@stanford.edu

Name: Riya Sinha
SUNet ID: riyasinh
Department of Biomedical Data Science
Stanford University
riyasinh@stanford.edu

Name: David Kuo MD
SUNet ID: dekuo
Department of Biomedical Data Science
Stanford University
dekuo@stanford.edu

Abstract

Acute myeloid leukemia (AML) has a poor prognosis, with a five-year overall survival rate of 30% [1]. Some patients' leukemia carries a particularly high risk of death or relapse, and these patients are usually recommended for bone marrow transplantation. However, it is often unclear which leukemia is at highest risk, and because bone marrow transplants are associated with significant morbidity and mortality, accurate risk stratification is crucial. Given the genetic and phenotypic heterogeneity of AML, this risk stratification can be challenging, and the standard system involves binning cases of AML using limited information about mutations and chromosomal abnormalities. We aimed to improve on this classification system by capturing both genotype and phenotype data from sequential bone marrow biopsy reports, which include the key elements to diagnose and characterize the AML and are used to follow disease. We used 8,131 bone marrow biopsy reports from 2,081 patients with AML at Stanford and GatorTron [2], a BERT model which had been previously trained on clinical and biomedical text, to predict one-year survival in cases of AML. We performed two main experiments: 1) comparing a patient-level model integrating a patient's sequential reports using an LSTM vs. a report-level model making predictions from individual reports only, and 2) exploring the impact of pre-training with supervised contrastive learning (SCL) [3] prior to fine-tuning vs. fine-tuning from the original GatorTron weights. Our best-performing model leveraged both the patient-level LSTM and SCL pretraining, achieving an accuracy of 0.702 and AUROC of 0.693, comparing favorably to the standard risk stratification criteria using a comparable Stanford dataset (best accuracy 0.609, AUROC 0.597).

1 Introduction

The current standard of care for AML risk prediction comes from the European LeukemiaNet (ELN) collaboration and assigns a patient to favorable, intermediate, or adverse risk categories based on mutations and chromosomal abnormalities [4] (Supplementary Table 1). These categories are then used to predict prognosis, which determines treatment and bone marrow transplant eligibility. Other models leveraging laboratory data [5] are also used; however all of these models struggle with limited expressivity, as they do not look at the full breadth of data available to an AML physician such as

the leukemia's phenotype, and do not incorporate patient data outside of a single time point. Recent literature has shown that the cancer phenotype is important in predicting treatment response in AML [6] and that integrating information over time improves prognostication in other cancers [7]. Thus, we set out to design a deep learning model that uses pathology reports to predict survival at one year by integrating reports over time for an individual patient. Each report contains phenotype and sometimes genotype information.

2 Related work

2.1 Pretrained Large Language Models for Biomedical and Clinical Text

In the past few years, multiple large language models have been trained using clinical text and medical literature. This includes ClinicalBERT, which was trained to predict 30-day hospital readmission [8] and was an inspiration for our project. However, ClinicalBERT was trained using data from an intensive care unit, and the vocabulary may have little relevance to AML prognostication. Other models have been trained using PubMed abstracts and available full-text academic papers, such as BioMegatron [9], or based on large amounts of general clinical text, such as GatorTron [2]. GatorTron was trained using >90 billion words of text (including >82 billion words of de-identified clinical text) from the University of Florida Medical Center. Given that GatorTron's training corpus specifically includes pathology reports and other clinical text, we chose to use this model for additional pre-training and fine-tuning.

Although there are multiple computational or expert-driven models to predict prognosis in AML and multiple large language models for clinical natural language processing, to our knowledge, no language models have thus far been developed to predict prognosis in AML.

2.2 Supervised Contrastive Learning

Self-supervised masked language modeling and contrastive learning have emerged as powerful pre-training techniques in natural language processing and computer vision. More recently, supervised contrastive learning was developed as an extension to self-supervised contrastive learning to leverage label information to generate higher quality positives and negatives from in-class and out-of-class examples compared to data augmentations from a given anchor example. This circumvents the need for explicit hard-negative mining since gradient contributions from hard positives and negatives are implicitly large while those for easy positives and negatives are small. In computer vision, supervised contrastive learning was found to achieve top-1 accuracy of 81.4% with ResNet-200 on the ImageNet dataset, an 0.8% improvement over the previous state-of-the-art, AutoAugment with cross entropy loss, using the same architecture. Similar improvements were observed with Vision Transformer ViT-B/16 and on CIFAR-100 and CIFAR-10 datasets [3]. Supervised contrastive learning has also been used for fine-tuning as an additional loss term to standard cross-entropy loss, as described in a recent study showing improvements over cross-entropy alone in language modeling using multiple datasets of the GLUE benchmark in few-shot learning settings [10].

3 Approach

3.1 Data Preprocessing

3.1.1 Data Truncation

We extracted and filtered pathology reports from an existing database through a pre-established collaboration with staff at the Stanford Cancer Institute Research Database (SCIRDB). Because the pathology reports were longer than GatorTron's 512 maximum token length, we parsed the reports so that the highest yield sections would fall within the 512-token limit. Specifically, we used the "Diagnosis" section, the "Comment" section, and the "Bone marrow biopsy and aspirate" or "Marrow" section of each report, concatenating these three sections in that that order (Diagnosis, Comment, Marrow). We used the HuggingFace `AutoTokenizer` to tokenize the reports and then truncated the reports to 512 tokens. Figure 1 depicts the number of words in the total concatenated report, Diagnosis section, and Comment section, respectively, where a "word" was separated by spaces, and below, we describe each section in more detail:

1. **Diagnosis:** summarizes the pathologist's main findings and concluding diagnosis
2. **Comment:** provides additional details of the pathologist's interpretation
3. **Marrow:** provides detailed descriptions of cells, bone marrow structure, any dysplasia, and manual counts of specific cell types

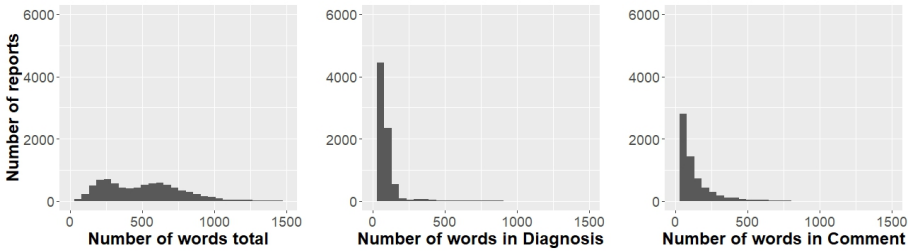


Figure 1: Distribution of the number of words (separated by spaces) in the total report, in the Diagnosis section, and in the Comment section.

3.1.2 Prompt Tuning

We wanted to incorporate information about the date a report was created relative to the patient's diagnosis date. To do this, we pre-pended each report with the text "Days from diagnosis: $DAYS". We hoped this additional context would be able to help the model understand the relationship of the report relative to the patient's timeline since diagnosis.

3.2 Baseline model

Our baseline comparator is the European LeukemiaNet (ELN) criteria [4], which is the standard risk stratification method for AML and is based entirely off of limited mutations and chromosomal abnormalities (e.g. no phenotypic data, Supplementary Table 1). Assignment of a subset of patients to ELN categories was previously done as part of on-going clinical research at Stanford. Supplementary Figure 1 depicts the ROC curve and AUC for the ELN criteria using a subset of this dataset (749 patients) for which the ELN data were available at diagnosis.

3.3 Pre-training Approaches

3.3.1 Domain-specific Pretrained Models

We chose GatorTron [2] as our base model to extend and fine-tune from because GatorTron's training corpus included a broad range of clinical text. We hypothesized that this would provide a good tokenizer for our text and initial parameters for our model compared to large language models trained only on PubMed abstracts or intensive care unit clinical notes.

3.3.2 Supervised Contrastive Learning

Additionally, we explored the use of supervised contrastive learning as an additional pre-training step before fine-tuning the model on our downstream binary classification task, hypothesizing that this would result in higher quality model weights and embeddings that could better separate positive and negative examples for faster and more effective downstream fine-tuning. We implemented our supervised contrastive loss function from Equation 2 in the original paper by Kholsa et al. [3]:

$$L_{out}^{sup} = \sum_{i \in I} L_{out,i}^{sup} = \sum_{i \in I} \frac{-1}{P(i)} \sum_{p \in P(i)} log \frac{exp(z_i \cdot z_p / \tau)}{\sum_{a \in A(i)} exp(z_i \cdot z_a / \tau)} \quad (1)$$

Here, i represents an anchor in the batch, $P(i)$ is the set of all other indices matching the label of i, and $A(i)$ is the set of all other anchors. z_i is the projection of the encoding of the data x_i. The temperature parameter, τ, is a hyperparameter that can affect the smoothness of the loss function. The loss function thus promotes the similarity of projections with the same label, while penalizing similar embeddings of different labeled samples. Of note, our implementation of supervised contrastive learning does not currently include additional data augmentations for positive and negative samples.

3.4 Patient-level Predictions

While our units of observation were individual reports from patient visits, our ultimate goal was to make predictions regarding the prognosis of patients, who may have multiple bone marrow biopsies and thus multiple pathology reports over time. Given that a series of reports of a single patient gives clinicians a more holistic view of a patient's health trajectory, we sought to design an architecture that could similarly utilize this property of the data.

Our general idea was to sequentially feed in each report from a patient into the BERT model, and apply some pooling layer to aggregate the report embeddings. The combined embedding could then be passed into a classifier for the final prediction.

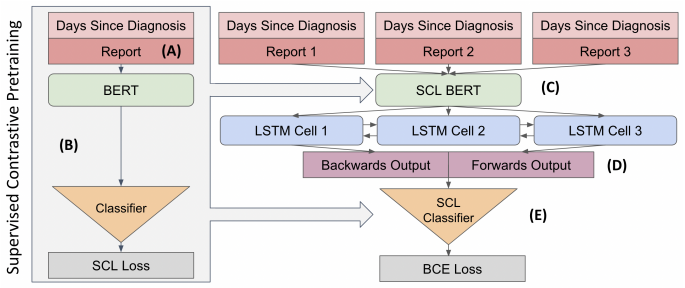


Figure 2: *Patient-level multi-report classification with SCL & bi-LSTM*: **(A)** The report is "prompt tuned" by prepending the number of days since diagnosis **(B)** SCL pretraining is performed **(C)** Report embeddings are generated using BERT **(D)** Report embeddings are fed into a bidirectional LSTM, and the final forward and backward LSTM hidden states are concatenated **(E)** The hidden states are passed to the final classification head.

We analyzed report embeddings as a sequence of events with a bidirectional Long Short-Term Memory (bi-LSTM) module, since it captures temporal relationships in both the forward and backward directions. The hidden state at the end of each direction encodes information from the whole sequence, and so these hidden states were concatenated and then classified as per the architecture in Figure 2.

3.5 Architecture

Our diverse set of approaches led us to implement a `ModularBertForSequenceClassification` loosely based off its HuggingFace counterpart. This version of the model is easily subclassable, featuring BERT embedding, pooling, classification, and loss functions that can be overridden, as illustrated in Figure 3. All models use the same `forward` method applying these modules in order. All classification heads include dropout with probability of 0.10.

4 Experiments

4.1 Data

Matthew Schwede is a hematologist/oncologist at Stanford who has been involved in building a database of patients with AML at Stanford by mining Stanford's electronic health record (EHR). Thus, the group has access to a large cohort (n = 2,240) of unique patients with AML who have bone marrow biopsy or peripheral blood flow cytometry reports (n = 10,415 unique reports). Pathology reports were only those that were listed as procedure reports in the EHR database and which had character strings of either "bone marrow biopsy" alone, or "peripheral blood" plus either "flow cytometry" or "immunophenotype". Other types of pathology reports (e.g. lumbar puncture pathology) were explicitly excluded. Death data were obtained throughs SCIRDB. After limiting reports to those that occurred before 365 days from diagnosis, there were 8,131 reports from 2,081 patients, 37% of whom died at one year. These data were split by patient to training:validation:test ratios of 80:10:10, without stratifying by class.

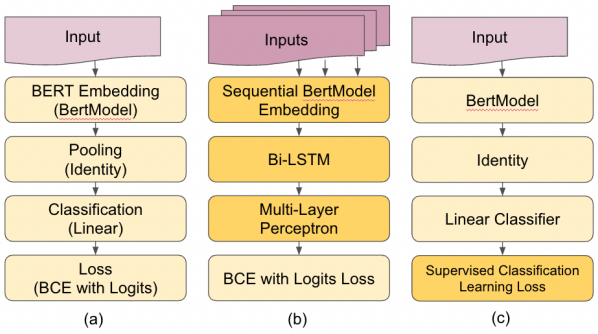


Figure 3: *Model architectures*: (a) The default architecture for `ModularBert` for classifying individual reports. Main text represents the module type, with the default in parentheses. Defaults are: normal BertModel, no pooling, linear classification and binary cross entropy (BCE) loss. (b) For patient classification, BERT is run sequentially on each report, Bi-LSTM pooling is applied, and a multi-layer perceptron (MLP) classifier is used. (c) Pretraining with supervised contrastive learning keeps all default components except the loss function.

4.1.1 Exploratory Data Analysis

Report and patient-level counts of negative vs. positive survival across training, validation and test sets are displayed in Table 1, showing a roughly 2:1 imbalance between the two classes. Additionally, histograms of the number of reports per patient and number of days since diagnosis for each report are shown in Figure 1: the majority of patients had less than 5 associated reports although a small minority of patients had over 10 reports, and reports were mostly clustered within the first 50 days of diagnosis.

Data level	Split	Positive Survival	Negative Survival
Report	Train	4050	2362
Report	Validation	543	298
Report	Test	590	288
Patient	Train	1010	654
Patient	Validation	590	288
Patient	Test	127	82

Table 1: Report-level and Patient-level survival at 1 year

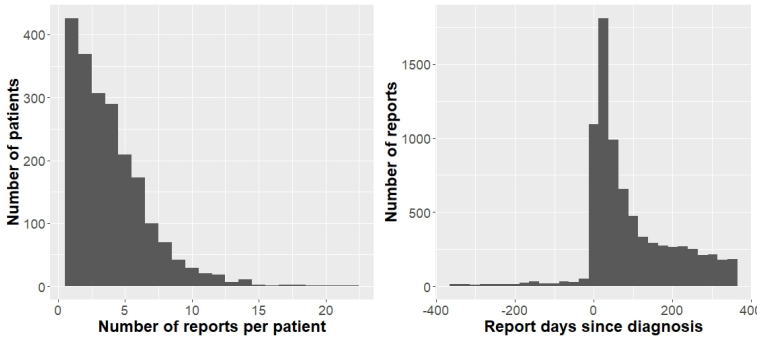


Figure 4: Number of reports per patient (left) and days between diagnosis and the report.

4.2 Evaluation method

For our binary classification task predicting survival at 1 year from diagnosis, we report accuracy, area under the receiver operating curve (ROC/AUC), precision (also known as positive predictive value), recall (also known as sensitivity), and F1 (the harmonic mean of precision and recall) as

evaluation metrics in order to gain as much insight into our model's performance especially in light of the class imbalance in our dataset.

4.3 Experimental details

Using GatorTron as our base model architecture, we experimented with a simple dropout layer followed by a linear layer vs. a multilayer perceptron of two linear layers separated by ReLU and dropout as our final classification head. We also explored the benefits of pre-training the initial GatorTron model weights with supervised contrastive learning with temperature hyperparameter $\tau = 1$ and $\tau = 0.1$ prior to full fine-tuning for binary classification. To account for class imbalance between our labels, we used the `pos_weight` parameter when calculating binary cross-entropy loss to prevent our model from fitting to just the majority label. For patient-level predictions, we also tried simpler pooling layers, such as mean pooling, to aggregate BERT embeddings of reports.

We used the AdamW optimizer [11] with learning rate 1e-5 and a linear learning rate scheduler with fractional warmup of 15%. Each model was trained for 50 epochs with the maximum tolerated batch size (4 for report-level data and 1 for patient-level data accommodating multiple reports over time) with gradient accumulation over 64 steps for an effective batch size of 256 and 64 for report-level data and patient-level data respectively. After encountering `CUDA:Out of Memory` errors when training on patient-level data with many associated reports, we implemented mixed precision training with `float16` to allow us to complete training runs. All training was completed on Google Cloud Platform (GCP) using NVIDIA 40GB A100 GPUs on `a2-highgpu-1g` Compute instances.

4.4 Results

We found that modeling at the patient level using a bi-directional LSTM to integrate multiple sequential reports improved performance over predicting survival from individual reports, consistent with previous findings that integrating more data during the patient's treatment course has been shown to improve survival prediction [7]. Supervised contrastive learning also improved model performance over fine-tuning from the original GatorTron model weights, in line with previous results [3, 10]. Indeed, our best model was the SCL-pretrained bi-LSTM-based patient-level model with an MLP classification head, which achieved an accuracy of 0.702 and an AUROC of 0.693 on the test set (Table 2).

Model	Acc	AUC	F1	Prec	Recall
GT/Reports	0.665	0.608	0.464	0.488	0.441
GT/Patients	0.688	0.614	0.444	0.634	0.342
SCL/Patients	0.702	0.693	0.617	0.581	0.658

Table 2: *Final Test Set Results*: Our best model used both pretraining with supervised contrastive learning and a bidirectional LSTM to incorporate patient-level data from reports over time.

5 Analysis

5.1 Tokenization and ELN Genetic Markers

We analyzed the tokenization of our reports to ensure that there was not a large number of `[UNK]` tokens given the specialized nature of our text. After excluding `[PAD]` tokens, single-character tokens, and NLTK stop words [12], we found that `[UNK]` was the 292nd most common token with 1408 instances counted across 8,131 reports. Examining the 30 most frequent tokens in our corpus after filtering, shown in Figure 5, showed that the GatorTron tokenizer was able to capture important medical terminology such as "blasts" as well as prefixes and suffixes such as "mye-" or "-ocytes". However, there were no ELN genetic markers (ex. "NPM1") in the GatorTron tokenizer vocabulary and only 1,241 total occurrences of ELN genetic markers across the 8,131 reports in our dataset (Supplementary table 2), suggesting that our model made predictions with minimal genetics data.

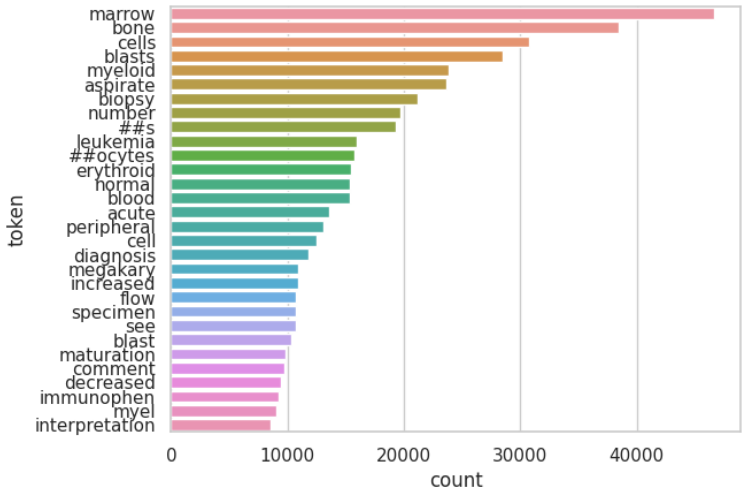


Figure 5: Top 30 most frequent tokens in dataset after filtering out stop words and 1 character tokens.

5.2 Supervised Contrastive Pretraining & BERT Learning

To understand the impact of our pretraining and finetuning of the models on the BERT embeddings, we used t-SNE plots (Figure 6) to look at the distribution of each class in a lower-dimensional space. These highlight the benefits of class separation using supervised contrastive learning. We experimented with different temperature parameters, and found that $\tau = 0.1$ provided the best results.

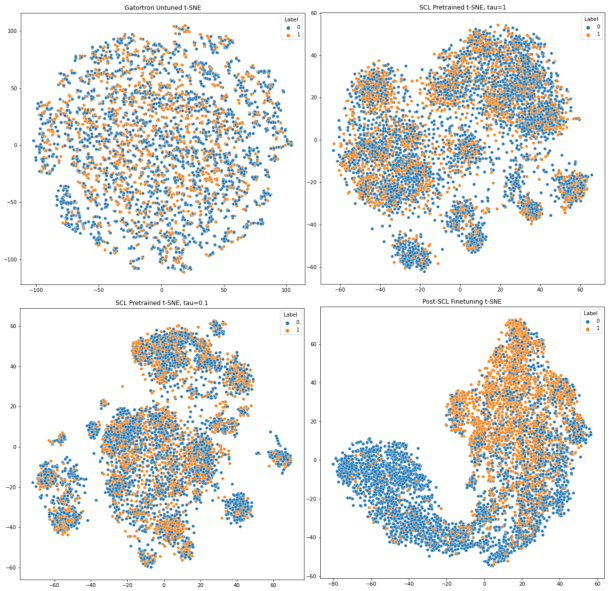


Figure 6: t-SNE plots of BERT pooled outputs at different model timepoints: raw GatorTron model (top-left), SCL pretraining with $\tau = 1$ (top-right), SCL pretraining with $\tau = 0.1$ (bottom-left), and post-finetuning the SCL pretrained model (bottom-right).

5.3 Ablation experiments

To understand whether select sections of the report were essential for the top model's performance, we performed ablation experiments where we removed report sections (e.g. Diagnosis) and then used our model to predict survival. Accuracy and AUROC were similar irrespective of which bone

marrow biopsy section was removed, although performance was worst when the Diagnosis section was removed, suggesting this section was of particular importance in classification.

Model	Acc	AUC	F1	Prec	Recall
No Diagnosis	0.682	0.681	0.606	0.561	0.658
No # Days	0.712	0.700	0.625	0.595	0.658
No Comment	0.692	0.691	0.619	0.565	0.684
No Marrow	0.688	0.681	0.606	0.562	0.658

Table 3: *Ablation experiment results*: Model performance with Diagnosis section, Comment section, Bone marrow biopsy and aspirate section, and number of days since diagnosis removed.

5.4 Manual Error Analysis

Correct vs. incorrect examples differed significantly (p=0.01) by the number of days from diagnosis, but not by diagnosis date, number of reports, report length, or report section (e.g. Diagnosis section) length (p>0.05, Supplementary Figure 2). Analyses using a subset (44%) of the dataset for which bone marrow blast percentages (% of cells that have morphology like leukemia cells) were previously extracted suggested that these percentages were lower in the correct cases (rank-sum p = 0.037), with roughly one half of reports having blast percentages < 5%, the cutoff for a morphologic remission. Additional manual review of 25 random correctly and incorrectly predicted reports was unremarkable.

This suggests that greater ambiguity regarding the clinical significance of the report (e.g. due to slightly higher blast percentages or more days from diagnosis) leads to worse performance. One report from an incorrectly classified patient highlighted the varying and evolving information used to classify AML: "While blasts are not increased by morphology in the bone marrow aspirate, the detection of cytogenetically abnormal cells supports the presence of a neoplastic clone." New types of data, new treatments, and new ways of interpreting data all affect modeling.

6 Future Work

Future directions include exploring better representations for modeling reports over time (for instance with a time positional encoding), using language models with longer sequence windows such as Clinical Longformer [13] to capture full pathology reports, exploring unsupervised pre-training via masked language modeling or unsupervised contrastive learning, adding data fusion (early or late) with other clinical data (e.g. ELN criteria, labs, mutations, treatment, other provider notes), and validating our findings with data from other institutions.

7 Conclusion

We demonstrated that leveraging serial pathology reports with large language models can produce more accurate predictions than the current standard risk stratification, despite not including key pieces of information in our model, such as mutations, cytogenetic abnormalities, or treatment. This is particularly remarkable because advances in AML have been sparse until recently. For many years (from the 1970s until 2017), there were only three therapies approved for AML, all of which involved very similar types of highly toxic chemotherapy [14]. This work suggests that a complex representation of serial pathology reports, which usually don't include the standard risk stratification variables, contain significant information that should be leveraged for better prognostication.

Overall, we learned a tremendous amount about the practical challenges of building a dataset; implementing, training, and debugging deep learning models; and critically understanding our data to analyze our model's performance. We also learned the importance of uninstalling `torch-xla` in Google Cloud when working with PyTorch (and coincidentally the viability of training LLMs on Apple Silicon), as well as the dangers of inconsistently saving and loading model checkpoints with and without `torch.nn.DataParallel`.

References

[1] Cancer Stat Facts: Leukemia — Acute Myeloid Leukemia (AML), 2022.

[2] Xi Yang, Aokun Chen, Nima PourNejatian, Hoo Chang Shin, Kaleb E Smith, Christopher Parisien, Colin Compas, Cheryl Martin, Anthony B Costa, Mona G Flores, et al. A large language model for electronic health records. *npj Digital Medicine*, 5(1):194, 2022.

[3] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning, 2020.

[4] Hartmut Döhner, Andrew H. Wei, Frederick R. Appelbaum, Charles Craddock, Courtney D. DiNardo, Hervé Dombret, Benjamin L. Ebert, Pierre Fenaux, Lucy A. Godley, Robert P. Hasserjian, Richard A. Larson, Ross L. Levine, Yasushi Miyazaki, Dietger Niederwieser, Gert Ossenkoppele, Christoph Röllig, Jorge Sierra, Eytan M. Stein, Martin S. Tallman, Hwei-Fang Tien, Jianxiang Wang, Agnieszka Wierzbowska, and Bob Löwenberg. Diagnosis and management of AML in adults: 2022 recommendations from an international expert panel on behalf of the ELN. *Blood*, 140(12):1345–1377, September 2022.

[5] Moritz Gerstung, Elli Papaemmanuil, Inigo Martincorena, Lars Bullinger, Verena I. Gaidzik, Peter Paschka, Michael Heuser, Felicitas Thol, Niccolo Bolli, Peter Ganly, Arnold Ganser, Ultan McDermott, Konstanze Döhner, Richard F. Schlenk, Hartmut Döhner, and Peter J. Campbell. Precision oncology for acute myeloid leukemia using a knowledge bank approach. *Nature Genetics*, 49(3):332–340, March 2017. Number: 3 Publisher: Nature Publishing Group.

[6] Shanshan Pei, Daniel A. Pollyea, Annika Gustafson, Brett M. Stevens, Mohammad Minhajuddin, Rui Fu, Kent A. Riemondy, Austin E. Gillen, Ryan M. Sheridan, Jihye Kim, James C. Costello, Maria L. Amaya, Anagha Inguva, Amanda Winters, Haobin Ye, Anna Krug, Courtney L. Jones, Biniam Adane, Nabilah Khan, Jessica Ponder, Jeffrey Schowinsky, Diana Abbott, Andrew Hammes, Jason R. Myers, John M. Ashton, Travis Nemkov, Angelo D'Alessandro, Jonathan A. Gutman, Haley E. Ramsey, Michael R. Savona, Clayton A. Smith, and Craig T. Jordan. Monocytic Subclones Confer Resistance to Venetoclax-Based Therapy in Patients with Acute Myeloid Leukemia. *Cancer Discovery*, 10(4):536–551, April 2020.

[7] David M. Kurtz, Mohammad S. Esfahani, Florian Scherer, Joanne Soo, Michael C. Jin, Chih Long Liu, Aaron M. Newman, Ulrich Dührsen, Andreas Hüttmann, Olivier Casasnovas, Jason R. Westin, Matthais Ritgen, Sebastian Böttcher, Anton W. Langerak, Mark Roschewski, Wyndham H. Wilson, Gianluca Gaidano, Davide Rossi, Jasmin Bahlo, Michael Hallek, Robert Tibshirani, Maximilian Diehn, and Ash A. Alizadeh. Dynamic Risk Profiling Using Serial Tumor Biomarkers for Personalized Outcome Prediction. *Cell*, 178(3):699–713.e19, July 2019.

[8] Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *CoRR*, abs/1904.05342, 2019.

[9] Hoo-Chang Shin, Yang Zhang, Evelina Bakhturina, Raul Puri, Mostofa Patwary, Mohammad Shoeybi, and Raghav Mani. BioMegatron: Larger biomedical domain language model. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4700–4706, Online, November 2020. Association for Computational Linguistics.

[10] Beliz Gunel, Jingfei Du, Alexis Conneau, and Ves Stoyanov. Supervised contrastive learning for pre-trained language model fine-tuning. *CoRR*, abs/2011.01403, 2020.

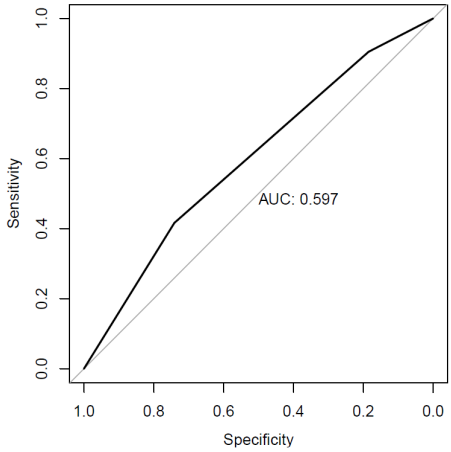
[11] Ilya Loshchilov and Frank Hutter. Fixing weight decay regularization in adam. *CoRR*, abs/1711.05101, 2017.

[12] Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.", 2009.

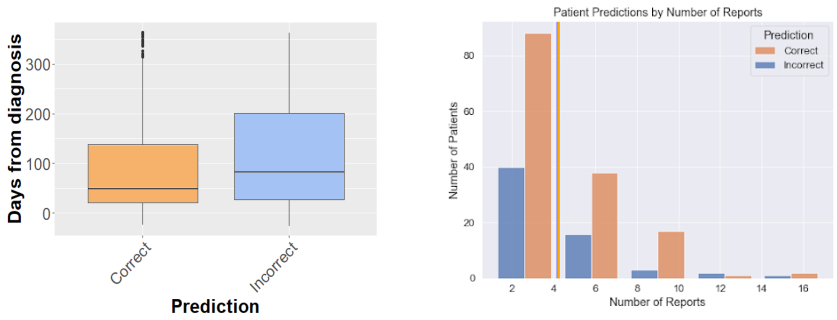
[13] Yikuan Li, Ramsey M. Wehbe, Faraz S. Ahmad, Hanyin Wang, and Yuan Luo. Clinical-longformer and clinical-bigbird: Transformers for long clinical sequences. *CoRR*, abs/2201.11838, 2022.

[14] Jenna L. Carter, Katie Hege, Jay Yang, Hasini A. Kalpage, Yongwei Su, Holly Edwards, Maik Hüttemann, Jeffrey W. Taub, and Yubin Ge. Targeting multiple signaling pathways: the new approach to acute myeloid leukemia therapy. *Signal Transduction and Targeted Therapy*, 5(1):288, December 2020.

8 Appendix



Supplementary Figure 1: ROC curve and AUC for predicting 1 year survival with the European LeukemiaNet Criteria (ELC)



Supplementary Figure 2: (left) Days from diagnosis vs. prediction correctness (right) and number of reports vs. prediction correctness. The colored lines represent the mean number of reports in the correctly and incorrectly classified patient sets.

European LeukemiaNet AML Risk Stratification 2022	
Risk	Features
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 TP53a

Supplementary Table 1: Risk stratification for acute myeloid leukemia per the European LeukemiaNet (ELN) recommendations in 2022.

ELN Genetic Marker	**Raw Data**	**Train/Val/Test Data**
RUNX1	847	96
CBFB	761	64
MYH11	125	10
NPM1	3220	335
FLT3	2952	305
MLLT3	63	17
KMT2A	245	30
DEK	49	00
NUP	49	0
BCR-ABL	388	41
KAT6A	0	0
CREBBP	13	6
GATA2	64	3
MECOM	94	4
ASXL1	119	6
BCOR	1308	81
EZH2	16	12
SF3B1	110	12
SRSF2	151	13
STAG2	72	6
U2AF1	10	0
ZRSR2	10	0
TP53	392	66

Supplementary Table 2: *ELN genetic marker counts*: There is a significant drop in text occurrences of ELN genetic markers between the raw dataset and our processed training, validation, and test datasets.