Agent-Based Feature Generation from Clinical Notes for Outcome Prediction

Jiayi Wang[1,*], Jacqueline Jil Vallon[1], Neil Panjwani[2], Xi Ling[2], Sushmita Vij[3], Sandy Srinivas[4], John Leppert[5,6,7], Mark K. Buyyounouski[2,†], Mohsen Bayati[2,8,9,†]

[1]Department of Management Science and Engineering, Stanford University School of Engineering, Stanford, CA
[2]Department of Radiation Oncology, Stanford University School of Medicine, Stanford, CA
[3]Graduate Business School Research Hub, Stanford University Graduate Business School, Stanford, CA
[4]Department of Medicine (Oncology), Stanford University School of Medicine, Stanford, CA
[5]Department of Medicine, Stanford University School of Medicine, Stanford, CA
[6]Department of Urology, Stanford University School of Medicine, Stanford, CA
[7]Veterans Affairs Palo Alto Health Care System, Palo Alto, CA
[8]Department of Electrical Engineering, Stanford University School of Engineering, Stanford, CA
[9]Operations, Information and Technology, Stanford University Graduate Business School, Stanford, CA

*Corresponding author. Email: jyw@stanford.edu
†Served as equally contributing co-senior authors

Abstract

Electronic health records (EHRs) contain rich unstructured clinical notes that could enhance predictive modeling, yet extracting meaningful features from these notes remains challenging. Current approaches range from labor-intensive manual clinician feature generation (CFG) to fully automated representational feature generation (RFG) that lack interpretability and clinical relevance. Here we introduce SNOW (Scalable Note-to-Outcome Workflow), a modular multi-agent system powered by large language models (LLMs) that autonomously generates structured clinical features from unstructured notes without human intervention. We evaluated SNOW against manual CFG, clinician-guided LLM approaches, and RFG methods for predicting 5-year prostate cancer recurrence in 147 patients from Stanford Healthcare. While manual CFG achieved the highest performance (AUC-ROC: 0.771 ± 0.036), SNOW matched this performance (0.761 ± 0.046) without requiring any clinical expertise, significantly outperforming both baseline features alone (0.691 ± 0.079) and all RFG approaches. The clinician-guided LLM method also performed well (0.732 ± 0.051) but still required expert input. SNOW's specialized agents handle feature discovery, extraction, validation, post-processing, and aggregation, creating interpretable features that capture complex clinical information typically accessible only through manual review. Our findings demonstrate that autonomous LLM systems can replicate expert-level feature engineering at scale, potentially transforming how clinical ML models leverage unstructured EHR data while maintaining the interpretability essential for clinical deployment.

1 Introduction

Over the past decade, access to electronic health record (EHR) data has become ubiquitous. As a consequence, the use of artificial intelligence (AI) to guide patient-level decision making in clinical practice has grown. One such application of AI is to create clinical risk prediction models [See, 2007, Bayati et al., 2014, Henry et al., 2015, Obermeyer and Emanuel, 2016, Wiens et al., 2016, Chen and Asch, 2017, Komorowski et al., 2018]. These models leverage patient data to predict clinically relevant outcomes and subsequently guide allocation of medical treatments. A key component in designing these models is generating the features (i.e., covariates) supplied to the model from structured data (e.g., laboratory values) and unstructured data (e.g., physician progress notes) [Xu et al., 2012, Zheng and Casari, 2018]. As a result, the methods in which features are derived from unstructured notes—and the extent to which clinical expertise is incorporated into this process—can profoundly affect both model performance and scalability [Verduijn et al., 2007, DeLisle et al., 2010, Zhao and Weng, 2011, Singh et al., 2015, Moskovitch et al., 2017, Roe et al., 2020].

At one end of the spectrum are methods involving no clinical input. Existing fully automated methods use pretrained embedding models or end-to-end neural architectures to automatically generate latent features from raw clinical text, typically without any human oversight or domain-specific guidance. We refer to this class of methods as **representational feature generation (RFG)**. These methods, including transformer-based language models such as ClinicalBERT [Alsentzer et al., 2019, Huang et al., 2020] and models fine-tuned on EHR corpora [Liu et al., 2018, Gehrmann et al., 2018, Dhrangadhariya et al., 2021], map clinical notes into dense vector representations, which are then used as inputs to downstream prediction models. The appeal of RFG lies in its scalability and minimal annotation burden. However, recent work has raised concerns about RFG's opacity, susceptibility to spurious correlations, and potential to amplify bias [Obermeyer and Emanuel, 2016, Gianfrancesco et al., 2018, Ghassemi and Nsoesie, 2022, Seyyed-Kalantari et al., 2021].

At the opposite end of the spectrum is patient-level **clinician feature generation (CFG)**, which involves extensive manual review of clinical notes by domain experts. In CFG, clinicians define clinically meaningful variables, extract them on a per-patient basis, and may even adapt definitions based on individual case context. Although CFG is labor-intensive and not scalable, it remains the gold standard in many settings—particularly when clinical nuance or context is critical. Manual abstraction efforts, such as those used in tumor registries or chart review-based studies [Nguyen et al., 2020, Fries et al., 2021], have demonstrated high-quality structured data generation, but at a substantial human cost.

Recent advances in large language models (LLMs) have dramatically enhanced a class of semi-automated feature generation methods that blend clinical input with automation, which we term **clinician-guided LLM feature generation (CLFG)**. Studies have shown that general-purpose LLMs like ChatGPT and DeepSeek R1, as well as domain-specific models such as ClinicalMamba and GatorTron, can accurately extract structured information from clinical notes when guided by expert prompts [Huang et al., 2024, et al., 2025b,a, 2024, Wei et al., 2021]. These structured features have been shown to improve the accuracy of clinical prediction models [Anderson et al., 2025, McInerney et al., 2023]. Together, this body of work illustrates the promise of LLMs in combining scalability with clinical relevance. However, existing approaches still depend on some level of clinician guidance to define target features, construct prompts, or interpret outputs.

In this work, we introduce **SNOW**, a novel agent-based **Scalable Note-to-Outcome Workflow** that autonomously generates structured features from unstructured clinical notes, without requiring any human intervention. Unlike prior approaches that depend on expert-specified variables or handcrafted prompts, SNOW leverages a modular architecture of specialized LLM agents designed to emulate clinical reasoning in an interpretable and automatable manner. We evaluate SNOW on the task of predicting 5-year prostate cancer recurrence using EHR data from Stanford Healthcare. Our findings show that methods involving clinician guidance—namely, clinician feature generation (CFG) and clinician-guided LLM feature generation (CLFG)—consistently outperform fully automated representational feature generation (RFG) in both predictive accuracy and interpretability. Importantly, SNOW matches the performance of manual CFG while achieving the scalability and efficiency of automation, effectively bridging the gap between clinical expertise and modern AI systems.

2 Methods

2.1 Collaborative Process

To begin this project, our team of clinicians and data scientists had extensive biweekly meetings for over a year, discussing the disease's natural history, treatment options, potential patient-specific features, and clinically relevant outcome measures. This process in part entailed reading physician progress notes as a team and understanding how a clinician utilizes these notes in practice, including what information is relevant and how it is typically summarized to make day-to-day clinical decisions.

This exploratory work and exchange of knowledge made it clear that, while physician progress notes include tremendous information that is leveraged in daily clinical practice to understand a patient's disease trajectory, some features cannot be extracted programmatically with a set of population-level rules (e.g., percentage core-length involvement with cancer per systematic core) and instead require extraction at a *per-patient* level with human reasoning.

To capture this complexity, we initially created a systematic process to apply per-patient clinical expertise in defining and extracting patient features from progress notes for manual clinician feature generation (CFG). This at a high level consists of two main steps: first, translating extensive knowledge learned through clinical practice and continued medical education into important domain-specific concepts required to understand the outcome of interest; and second, converting the concepts into patient-level features that can be manually retrieved from unstructured data sources with further assistance of clinical expertise. We then subsequently complete the necessary time-intensive process using data from patients who were treated for prostate cancer at the Stanford Cancer Institute between 2005-2015. This study was reviewed and approved by the institutional review board (IRB) of Stanford University (IRB-49456) and informed consent was obtained from all participants.

Motivated by the need for a scalable alternative, we leveraged our experience in CFG to design SNOW, an agent-based Scalable Note-to-Outcome Workflow. SNOW consists of a sequence of specialized LLM agents, each emulating a core subtask in the CFG pipeline—such as feature discovery, extraction, validation, post-processing, and aggregation. Unlike prior LLM-based approaches that require detailed instructions and pre-specified target features, SNOW autonomously coordinates these tasks, enabling fully automated, interpretable feature generation from clinical text.

All LLM operations in this study were performed via the Secure GPT API provided by Stanford Health Care and Stanford School of Medicine. This API is approved for use with sensitive data, including Protected Health Information (PHI) and Personally Identifiable Information (PII), and adheres to institutional data security and privacy standards. Using this secure API ensured that all LLM-based feature extraction was conducted in compliance with HIPAA regulations and appropriate ethical guidelines for handling clinical data.

For each patient, we extract data from the Electronic Health Records (EHR). Available patient data includes structured fields (e.g., laboratory values) and unstructured physician progress notes (e.g., biopsy reports and clinical notes). We apply minimal clinical expertise to create a set of baseline features from the structured data sources. For the unstructured notes, we run the agentic system to generate a set of structured features. We also identify and extract features through our patient-level CFG protocol. To contrast the performance to an RFG method, we apply automated natural language processing (NLP) methods to the progress notes for feature extraction, and subsequently compare the performance of machine learning models trained and tested on varying feature sets to establish our conclusions.

2.2 Patient Cohort

Inclusion criteria for the patient cohort ($n = 147$) are: individuals who have a prostate-specific antigen (PSA) lab value available $>$ 5 years after treatment, have a pre-treatment biopsy report with the 12 systematic cores recorded ([right/left] [apex/mid/base] [lateral/medial]), and received a prostatectomy or radiation therapy as their first treatment and had no subsequent treatment within five years. The first criterion ensures we do not misclassify patients as not having cancer recurrence due to loss to follow-up, as the outcome of cancer recurrence is computed from PSA lab values. The second criterion constrains the study to patients who have sufficient information in a pre-treatment biopsy report to capture a representative description of their disease state. We perform a sensitivity analysis that captures the timing of this biopsy report that contains the 12 systematic cores and show results are qualitatively unchanged. Finally, the third criterion is necessary because we are interested in whether a patient's cancer recurs after their first treatment (prostatectomy or radiation therapy). This last criterion excludes an additional 21 patients who received a treatment that entailed a prostatectomy and then radiation therapy within a timespan of five years. In this case, radiation therapy is typically an adaptive treatment plan based on patient-specific risk factors for failure, and it can be difficult to extrapolate the counterfactual outcome of whether the patient would have experienced cancer recurrence had the patient only undergone their first treatment, posing difficulties in computing an unbiased estimate of the outcome for this cohort. While the above inclusion criteria may create inadvertent biases in the data, we expect such biases (if present) to not impact our comparison of different feature generation approaches and the main conclusions of the chapter. Table 1 shows key summary statistics for our final patient cohort.

2.3 Outcome

All models in our analyses predict the probability of biological failure (BF) within five years after the end of treatment signaling cancer recurrence. Biochemical failure is the gold-standard for defining recurrent prostate cancer and an important endpoint because it prompts an evaluation for local recurrence and metastatic disease, and in some cases additional therapy. For patients who receive radiation therapy, a patient is classified as BF if they have a PSA level post-treatment that is at least 2 ng/mL greater than their post-treatment nadir [Roach et al., 2006]. For patients who undergo a prostatectomy, they are classified as BF if they have a PSA level of 0.4 ng/mL or above with the subsequent PSA level increasing after the end of treatment [Stephenson et al., 2006]. Given there is variability in the latter definition, we

Table 1: Summary statistics of patient cohort.

Continuous Features	**Mean (SD)**	**Median**
Age at Treatment	68.5 (8.4)	68.7
Charlson Comorbidity Index	5.2 (2.0)	5.0
Maximum Pre-treatment PSA	10.8 (21.5)	6.2
Percent Positive Regions	40.8 (26.5)	33.3
Binary/Categorical Features	**Number (%)**	
Patients with biological failure	11 (7.5)	
Race = White	110 (74.8)	
Staging		
t1	75 (51.0)	
t2	47 (32.0)	
t3	12 (8.2)	
Grade Group of Max Gleason Score		
1	37 (25.2)	
2	36 (24.5)	
3	31 (21.1)	
4	17 (11.6)	
5	26 (17.7)	
Total number of patients	**147**	

complete a sensitivity analysis on the results in which we classify a patient who got a prostatectomy as BF if they have a PSA level post-prostatectomy of 0.2 ng/mL or above with the subsequent PSA level also being 0.2 ng/mL or above [Association, 2013].

2.4 Baseline Features

Baseline features include features from structured data sources that require minimal data analysis to compute with the output from the EHR. These include demographic and socioeconomic features, the maximum pre-treatment PSA level, and the Charlson Comorbidity Index [Glasheen et al., 2019]. We include the features of race, ethnicity, and language in our study, but perform a sensitivity analysis on the results excluding these three features to ensure that the outcome prediction machine learning models are not using these features in any biased fashion. For the purpose of our study, the baseline features require minimal clinical expertise to identify, either because they are readily available in the EHR (e.g., PSA values) or because there has been extensive research already conducted on how to compute them providing us with a template to follow (e.g., Charlson Comorbidity Index).

2.5 SNOW, an agent-based Scalable Note-to-Outcome Workflow

To automate the construction of clinically relevant, structured features from unstructured clinical notes, we designed a modular agentic system composed of specialized agents. Each agent is responsible for a specific subtask in the feature generation pipeline, operating in sequence to transition from feature discovery and extraction to post-processing and aggregation. This section outlines the responsibilities and behavior of each agent in the system.

The Feature Discovery Agent initiates the pipeline by scanning a corpus of clinical notes to propose a set of structured variables that are clinically meaningful and suitable for outcome prediction modeling. It excludes any features already available from structured data sources and considers the clinical context and prediction target. This agent identifies whether features are specific to subgroups—such as anatomical regions, timepoints, or other repeated contexts—or require aggregation across multiple measurements. It provides descriptive guidance and candidate extraction logic accordingly.

Next, the Feature Extraction Agent processes individual clinical notes and attempts to extract values for each proposed feature. It applies instructions from the discovery agent if this is the first extraction, or the updated guidance from the validation agent in cases of re-extraction, to parse relevant information from the note text. The outputs include raw or categorical values aligned with the proposed features.

Following extraction, the Feature Validation Agent performs quality control on the extracted values. It reviews a sample of clinical notes alongside the extracted values to assess their accuracy, completeness, and consistency. Based on this assessment, it decides whether to proceed with the feature, remove it, re-extract with revised instructions, or apply post-processing. This initiates a validation loop: if a feature is re-extracted or post-processed, it is returned to the validation agent for reevaluation. This iterative loop continues until the feature receives a final decision of either proceed or remove.

For features requiring transformation rather than re-extraction, the Post-Processing Agent applies additional logic such as normalization, relabeling, or binning. It follows specific transformation instructions issued by the validation step and revisits the source notes when necessary to finalize cleaned feature values. Post-processed features are returned to the Feature Validation Agent for reassessment, maintaining the integrity of the validation loop.

If a feature is defined as an aggregate of other base features—such as the mean or maximum across anatomical regions—the Aggregation Code Generator creates executable Python code to compute this value. It ensures the final aggregated feature handles missing values gracefully and adheres to the intended computation logic.

These agents interact through a modular and adaptive workflow. The process begins with the Feature Discovery Agent identifying candidate variables. Features identified as non-aggregated move directly to the extraction stage, while aggregated features are deferred until their base components are available. All extracted features pass through validation and, if required, undergo post-processing or re-extraction. Importantly, any revised output reenters the validation loop to ensure robustness. Finally, aggregated features are computed using the generated code.

SNOW supports autonomous, interpretable, and iterative feature generation. It enables scalable structured data generation for clinical modeling with minimal human oversight.

2.6 Patient-Level Clinician Feature Generation (CFG)

To generate the patient-level clinician features, over the span of a year from the end of 2020 to 2021, our team of oncologists (M.K.B., N.P., S.S.) and data scientists (J.J.V., M.B., S.V., X.L.) at the Stanford Cancer Institute collaborated to curate clinically relevant features from the medical record. The overarching process, as illustrated in Figure 1, leverages clinical expertise to evaluate a patient's likelihood of prostate cancer recurrence to define the clinical concepts and their associated features that comprise the curated dataset. For example, one concept is that tumor volume in the biopsy specimen is prognostic and features representative of tumor volume include the percentage of positive regions (i.e., first grouping cores by the region of the prostate from which they were taken and then determining whether a region had any positive cores) and the percentage core-length involvement with cancer per systematic core. All progress notes examined by the team in this step were pre-treatment, ensuring that information of the outcome variable was unknown during this process. Our team of oncologists also identified clinically relevant summary statistics that represent information both used in practice and taught to be of importance in clinical training and must therefore be generated as additional features. The result of this step is an acquired list of features that are both of clinical importance in predicting prostate cancer recurrence and available in the data.

Then, to identify and extract these named clinician features from notes, while a handful of them, such as the clinical T stage [Buyyounouski et al., 2017] and maximum Gleason score, can be extracted on a cohort level leveraging regular expressions, the majority of features from the biopsy reports require manual parsing on a per-patient level because of the unstructured and unstandardized nature of the reports. This retrieval step requires even further clinical input, this time from pathologists who write the biopsy reports, to correctly understand the variable wording in the texts and to extract the features correctly. Traits of our data that contribute to the required per-patient generation of the clinician features from notes include: 1) different naming conventions for regions and different number of regions sampled; 2) distinct characters (or none at all) to separate information for each region; 3) diverse reporting of results for a region (some reports combine cores per region if they have the same results, while others keep each core strictly separate); 4) variable structures, especially for reports summarizing slides submitted from an outside institution; 5) inconsistent notation for the percent of the core involved (for example, some

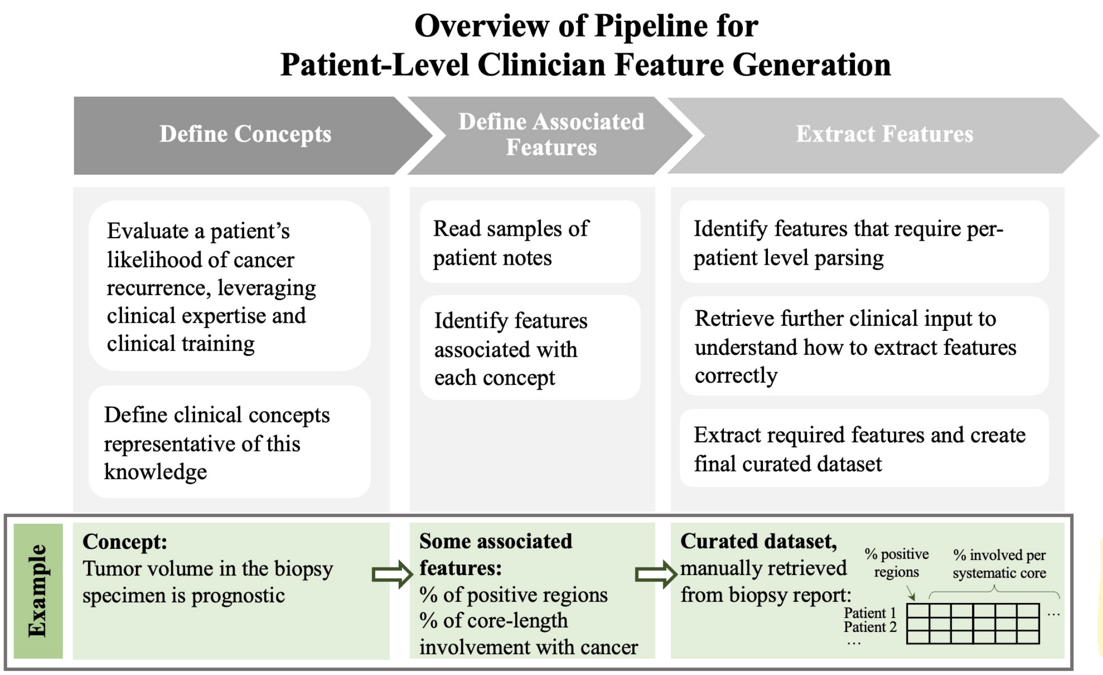

Overview of Pipeline for
Patient-Level Clinician Feature Generation
Define Concepts
Define Associated Features
Extract Features
Evaluate a patient's likelihood of cancer recurrence, leveraging clinical expertise and clinical training
Define clinical concepts representative of this knowledge
Read samples of patient notes
Identify features associated with each concept
Identify features that require per-patient level parsing
Retrieve further clinical input to understand how to extract features correctly
Extract required features and create final curated dataset
Example
Concept:
Tumor volume in the biopsy specimen is prognostic
Some associated features:
% of positive regions
% of core-length involvement with cancer
Curated dataset, manually retrieved from biopsy report:
% positive regions
% involved per systematic core
Patient 1
Patient 2
…

Figure 1: Overview of pipeline used to generate CFG features.

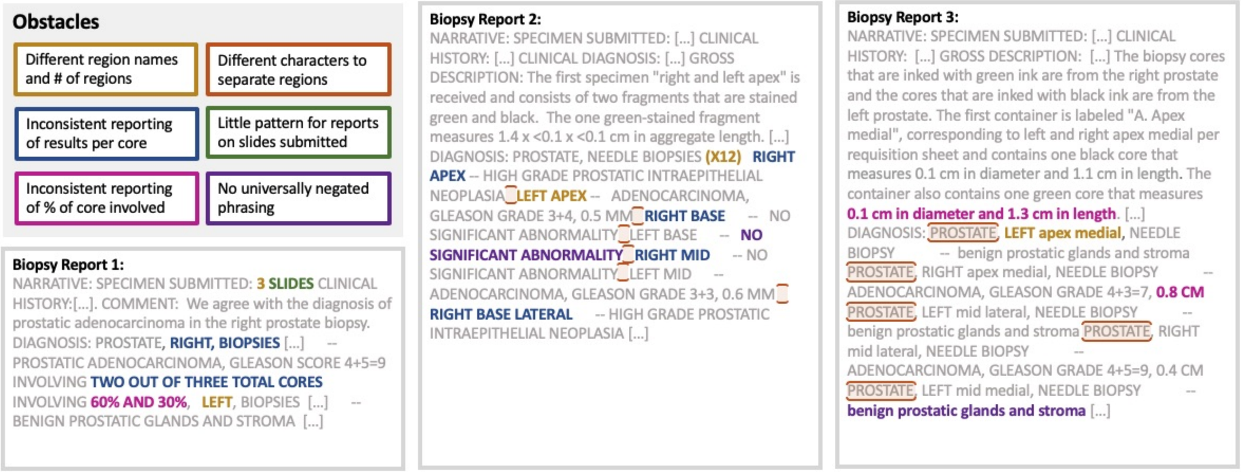
texts only state the length of the core involved in the diagnosis section, which must subsequently be divided by the length of the core found in the denser part of the text to compute the percent of the core involved, and this is made more difficult to do systematically with reports varying in the units used, including cm and mm); and 6) irregular wording for negated phrases such as "no adenocarcinoma" and "no significant abnormalities". Figure 2 shows examples of each of these obstacles in patient texts to illustrate that it is very hard to create a set of rules that could be programmatically implemented to extract these features at large (note: values in the figure have been changed for patient privacy reasons, but the overall layouts of the texts represent real patient texts).

Ultimately, the patient-level clinician features provide a very detailed depiction of each patient's tumor that expands beyond the commonly used higher-level features of tumor staging and Gleason grade. The patient-level curation enables the identification of features per systematic core and the calculation of features that may require clinical proxies to be defined (e.g., percent Gleason pattern 4/5). We denote this entire set of features using extensive patient-level clinical expertise as "CFG features".

2.7 Representational Feature Generation (RFG)

To retrieve features from progress notes using RFG, we apply different natural language processing (NLP) methods. For each patient, we create a concatenated progress note text, by combining the pre-treatment clinical note that has clinical T stage information present closest to the start date of treatment, identified using regular expressions, with the pre-treatment biopsy report from which we retrieve the subset of CFG features per systematic core (i.e., grade group, percent involved, and percent Gleason pattern 4/5 per systematic core). We then apply six different NLP methods to this concatenated text to generate features: Bag-of-Words (BoW Classic), Bag-of-Words Term Frequency-Inverse Document Frequency (BoW TF-IDF), Bidirectional Encoder Representations from Transformers (BERT), DistilBERT, ClinBERT, and Longformer.

BoW is a simple and interpretable NLP method. For all BoW-based models, we first preprocess the data by converting all text to lowercase, removing common English stopwords (using the `nltk` package in Python), and lemmatizing each word (i.e., reducing each word to its root form). We then build a corpus from the notes of the 147 patients. To capture varying levels of local context, we experimented with different n-gram configurations, including unigram–bigram (BoW 2-gram), unigram–trigram (BoW 3-gram), and unigram–4-gram (BoW 4-gram) models. BoW Classic refers to the version in which the feature value for each n-gram is the raw frequency count of its appearance in a clinical note. BoW TF-IDF



Obstacles

Different region names and # of regions

Different characters to separate regions

Inconsistent reporting of results per core

Little pattern for reports on slides submitted

Inconsistent reporting of % of core involved

No universally negated phrasing

Biopsy Report 1:
NARRATIVE: SPECIMEN SUBMITTED: **3 SLIDES** CLINICAL HISTORY:[...]. COMMENT: We agree with the diagnosis of prostatic adenocarcinoma in the right prostate biopsy. DIAGNOSIS: PROSTATE, **RIGHT, BIOPSIES** [...] -- PROSTATIC ADENOCARCINOMA, GLEASON SCORE 4+5=9 INVOLVING **TWO OUT OF THREE TOTAL CORES** INVOLVING **60% AND 30%**, **LEFT**, BIOPSIES [...] -- BENIGN PROSTATIC GLANDS AND STROMA [...]

Biopsy Report 2:
NARRATIVE: SPECIMEN SUBMITTED: [...] CLINICAL HISTORY: [...] CLINICAL DIAGNOSIS: [...] GROSS DESCRIPTION: The first specimen "right and left apex" is received and consists of two fragments that are stained green and black. The one green-stained fragment measures 1.4 x <0.1 x <0.1 cm in aggregate length. [...] DIAGNOSIS: PROSTATE, NEEDLE BIOPSIES **(X12)** **RIGHT APEX** -- HIGH GRADE PROSTATIC INTRAEPITHELIAL NEOPLASIA **LEFT APEX** -- ADENOCARCINOMA, GLEASON GRADE 3+4, 0.5 MM **RIGHT BASE** -- NO SIGNIFICANT ABNORMALITY LEFT BASE -- **NO SIGNIFICANT ABNORMALITY** **RIGHT MID** -- NO SIGNIFICANT ABNORMALITY LEFT MID -- ADENOCARCINOMA, GLEASON GRADE 3+3, 0.6 MM **RIGHT BASE LATERAL** -- HIGH GRADE PROSTATIC INTRAEPITHELIAL NEOPLASIA [...]

Biopsy Report 3:
NARRATIVE: SPECIMEN SUBMITTED: [...] CLINICAL HISTORY: [...] GROSS DESCRIPTION: [...] The biopsy cores that are inked with green ink are from the right prostate and the cores that are inked with black ink are from the left prostate. The first container is labeled "A. Apex medial", corresponding to left and right apex medial per requisition sheet and contains one black core that measures 0.1 cm in diameter and 1.1 cm in length. The container also contains one green core that measures **0.1 cm in diameter and 1.3 cm in length**. [...] DIAGNOSIS: PROSTATE, **LEFT apex medial**, NEEDLE BIOPSY -- benign prostatic glands and stroma PROSTATE, RIGHT apex medial, NEEDLE BIOPSY -- ADENOCARCINOMA, GLEASON GRADE 4+3=7, **0.8 CM** PROSTATE, LEFT mid lateral, NEEDLE BIOPSY -- benign prostatic glands and stroma PROSTATE, RIGHT mid lateral, NEEDLE BIOPSY -- ADENOCARCINOMA, GLEASON GRADE 4+5=9, 0.4 CM PROSTATE, LEFT mid medial, NEEDLE BIOPSY -- **benign prostatic glands and stroma** [...]

Figure 2: Examples of obstacles in automating the extraction of CFG features.

is an advanced variant that uses the same n-gram extraction and preprocessing pipeline but reweights n-gram frequencies using term frequency–inverse document frequency to downweight commonly occurring terms that are less informative across the corpus.

BERT, DistilBERT, ClinBERT, and Longformer, on the other hand, are deep-learning powered methods that have shown improved performance in numerous applications [Alsentzer et al., 2019, Hahn and Oleynik, 2020, Si et al., 2019]. BERT [Devlin et al., 2019], the forerunner of these methods, leverages transfer learning from attention-based transformers, enabling it to understand the complex and evolving meaning of a word in a text in relation to all surrounding words, rather than strictly focusing on the interpretation of a word from reading text left to right. This model was pre-trained on the vast English Wikipedia and the BooksCorpus. DistilBERT [Sanh et al., 2020], a variant of BERT, is considered to be faster to pre-train, while still being able to retain most of the understanding capabilities. Similarly, ClinBERT [Alsentzer et al., 2019] uses the same conceptual framework as BERT but is instead tuned specifically to the medical context by training on clinical notes in the MIMIC database. Alternatively, Longformer [Beltagy et al., 2020] is modified from BERT such that it can process longer text sequences. For these methods, we split the raw data into chunks of equal size that each method can process (chunks of size ∼512 tokens for BERT, DistilBERT, and ClinBERT and of size ∼4096 tokens for Longformer), apply the method onto each chunk, and concatenate the features.

To further enhance the contextual understanding relevant to our domain, we developed a fine-tuned embedding model based on the open-sourced Mistral-7B-v0.3 large language model [Jiang et al., 2023]. The fine-tuning dataset consisted of 16,872 pathology reports, from which all notes corresponding to the study cohort were carefully excluded. Fine-tuning was performed using a causal language modeling objective, where the model was trained autoregressively to predict the next token given all preceding tokens in the sequence. This approach enables the model to learn contextual dependencies relevant for generating semantically rich embeddings. This model was fine-tuned using Low-Rank Adaptation (LoRA) [Hu et al., 2021] with 8-bit quantization to optimize computational efficiency while retaining performance. We configured LoRA with a rank of 64 and a scaling factor of 128, targeting the query, key, value, and embedding layers for adaptation, with a dropout rate of 0.1 and no bias training. Training was conducted over 5 epochs with a batch size of 2 per device and gradient accumulation over 8 steps, yielding an effective batch size of 16. All training was performed on a single NVIDIA A100 GPU. After fine-tuning, we computed the embeddings by applying both mean pooling across all token hidden states and by extracting the hidden state corresponding to the end-of-sequence (EOS) token.

In addition to our fine-tuned Mistral model, we also evaluated embeddings generated from OpenAI's text embedding models, using both the `text-embedding-3-small` ("small") and `text-embedding-3-large` ("large") variants. This allowed us to assess whether general-purpose, proprietary embedding models could extract clinically useful signal in comparison to both our domain-specific models and other RFG approaches.

To ensure consistency across RFG methods and to mitigate overfitting in our relatively small dataset, we applied singular value decomposition (SVD) to reduce the dimensionality of the feature matrix for each method.

Table 2: Summary of Representational Feature Generation (RFG) Approaches

Method	Configurations	Number
BoW Classic	2-gram, 3-gram, 4-gram	3
BoW TF-IDF	2-gram, 3-gram, 4-gram	3
BERT Variants	BERT, DistilBERT, ClinBERT, Longformer	4
Fine-tuned Mistral-7B-v0.3	Mean pooling, EOS pooling	2
OpenAI Embedding Models	`text-embedding-3-small`, `text-embedding-3-large`	2
Total number of RFG approaches		**14**

2.8 Clinician-Guided LLM Feature Generation (CLFG)

We also explore using LLMs to generate patient-level features through expert-guided prompts. In each prompt, we ask the LLM to identify and extract the same set of clinician-defined features used in CFG. To support accurate extraction, we include detailed instructions based on our experience manually processing these features. For instance, when core-length involvement is reported as a raw measurement (e.g., in millimeters) rather than a percentage, we instruct the model to perform the necessary division. The LLM processes each note individually, extracting all relevant features one note at a time. For features that depend on other extracted values, such as percent Gleason pattern 4/5 or maximum Gleason score which are derived from individual Gleason scores, we apply post-processing code that follows clinical rules based on expert knowledge and standard medical guidelines.

2.9 Outcome Prediction Machine Learning Model

To obtain an unbiased estimate of model performance, we use nested cross-validation. Specifically, we implement a 3-fold outer cross-validation to evaluate generalization performance. For each outer fold, the remaining two-thirds of the data—the outer training set—are used to perform a 3-fold inner cross-validation to select the optimal penalty parameter λ based on area under the receiver operating characteristic curve (AUC-ROC) as the performance metric. After hyperparameter tuning within the inner folds, we retrain the model on the full outer training set using the best λ, and evaluate its performance on the held-out outer fold. This procedure ensures that the outer test data remain entirely separate from both model fitting and hyperparameter selection, thus providing an unbiased estimate of model performance.

Since it is known that formal statistical tests comparing AUC-ROC between different models are unreliable with small samples [Newcombe, 2006, Feng et al., 2017], we repeat the entire nested cross-validation procedure 50 times, each time using a different random seed to generate the outer and inner cross-validation splits. We then compare the distribution of the AUC-ROC of the models across the 50 seeds.

We used logistic regression with hyperparameter tuning over the regularization parameter $\lambda \in [1, 10^9]$. Both L1 and L2 penalties were considered. Missing values were handled using mean imputation. We also experimented with alternative imputation methods including singular value decomposition (SVD) and multivariate imputation by chained equations (MICE), but found they did not have a substantial impact on the results. Therefore, we report results using mean imputation for simplicity. A comprehensive sensitivity analysis of different imputation methods will be included in a future version of this work.

3 Results

Figure 3 presents the distribution of AUC-ROC scores across 50 iterations of nested cross-validations on the main outcome prediction model, comparing four feature sets: Baseline, Baseline + CLFG, Baseline + SNOW, and Baseline + CFG.

The 'Baseline + CFG' model achieves a substantially higher mean AUC (0.771 ± 0.036) compared to the 'Baseline' model (0.691 ± 0.079), demonstrating that clinician-guided patient-level feature generation (CFG) from progress notes significantly enhances predictive performance. This finding aligns with prior work on leveraging unstructured clinical notes for predictive modeling [Hsu et al., 2020, Liu et al., 2018, Ford et al., 2016].

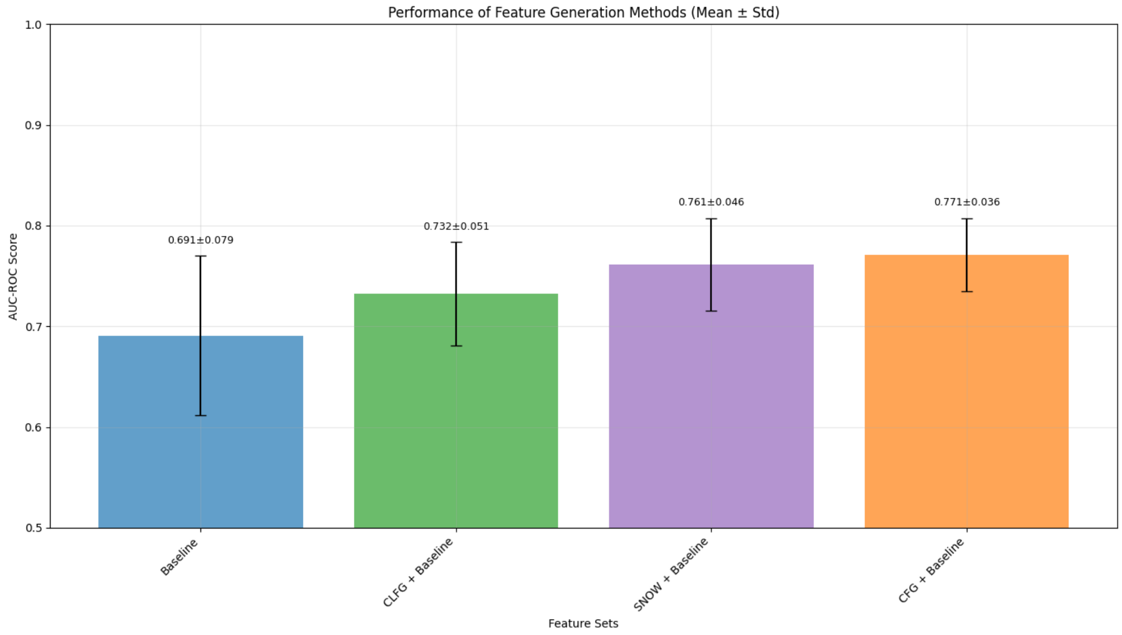


Figure 3: Comparison of the distribution of the AUC-ROC for different feature generation methods

The ‘Baseline + CLFG’ model (0.732 ± 0.051), which leverages large language models guided by clinician-written instructions, nearly matches the performance of the CFG model, suggesting that LLMs can effectively implement expert knowledge in feature generation from unstructured notes.

Notably, the ‘Baseline + SNOW’ model (0.761 ± 0.046), which fully automates the feature generation process using a multi-agent LLM system without human intervention, achieves performance on par with the manual CFG process. This result demonstrates that our agentic LLM system can effectively replace labor-intensive expert-driven CFG, enabling scalable and accurate feature generation for clinical prediction tasks. The feature specifications generated by SNOW are provided in Appendix A.

We comprehensively evaluated all 14 RFG approaches described in Table 2, implementing multiple strategies for combining RFG embeddings with baseline features. These strategies included applying singular value decomposition (SVD) for dimensionality reduction and combining the reduced embeddings with baseline features in various configurations. Despite this extensive evaluation, none of the ’RFG + baseline’ approaches outperformed the baseline features alone. These results suggest that while RFG methods likely introduce new signal, they also increase the dimensionality of the feature space and may complicate model training in a small-sample setting. This added complexity makes the model harder to optimize effectively. Because they did not improve model performance, we excluded them from Figure 3. In the next version of this paper, we plan to conduct these experiments on a larger dataset, where the benefits of high-dimensional RFG approaches may be more fully realized.

4 Discussion

Our study demonstrates that while patient-level clinician feature generation (CFG) substantially enhances prediction of 5-year prostate cancer recurrence, a fully automated system, SNOW, can achieve comparable performance. This system, composed of specialized agents that emulate clinical reasoning in a modular and interpretable fashion, eliminates the need for manual curation while preserving the clinical richness traditionally only achievable through expert review.

Historically, CFG has remained the gold standard for leveraging nuanced information from unstructured clinical notes. It relies on in-depth domain knowledge to identify and extract predictive features on a per-patient basis. Although this process delivers strong predictive performance, it is not scalable due to the time and expertise required. In contrast, representational feature generation (RFG) methods, such as embeddings from pretrained language models, offer scalability but often lack interpretability, and as we show, they fail to capture the same predictive signal.

SNOW introduces a new path: a fully autonomous agentic system that not only automates the feature generation process but also replicates the interpretability and performance of CFG. This system leverages domain-specialized LLM agents for feature discovery, extraction, validation, post-processing, and aggregation, ensuring that feature definitions are clinically grounded and rigorously vetted without human intervention. Our results show that the 'Baseline + SNOW' model achieves performance statistically comparable to the 'Baseline + CFG' model, significantly outperforming both RFG and baseline methods. This marks the first demonstration—within the context of prostate cancer recurrence prediction—of a fully automated system matching expert-driven performance using progress notes.

While RFG methods offer scalability and require minimal manual effort, our results show that they do not provide additional predictive value beyond baseline features in our current setting. The combination of high-dimensional embeddings, lengthy clinical notes, and a relatively small patient cohort likely limits the ability of RFG to extract meaningful signal. In contrast, the selective and targeted feature engineering performed by both CFG and SNOW centers on clinically relevant abstractions that are closely aligned with the prediction task, which appears to be critical for performance. In a future version, we plan to evaluate our approach on a larger dataset, where the strengths of RFG methods may be more effectively realized.

We also examine a hybrid setting where LLMs are guided by expert-written prompts to extract a predefined set of clinician features (referred to as CLFG in this paper). While this semi-automated approach performs nearly as well as manual CFG, it still relies on domain expertise to define the target features and to engineer prompts. SNOW removes this dependency entirely, discovering and validating features autonomously.

These findings highlight a fundamental shift in how clinical knowledge can be encoded and operationalized at scale. Rather than relying on static expert-defined feature sets or opaque embedding vectors, our agentic system dynamically generates structured features that are tailored to the specific clinical context. This capability enables not only scalability but also generalizability: because the system operates at the level of interpretable features and iterative validation, it may be more robust to institutional differences in documentation practices.

Looking ahead, our work suggests that LLMs can serve as intelligent collaborators in clinical ML pipelines—automating labor-intensive tasks, reducing annotation bottlenecks, and facilitating large-scale studies with interpretable and clinically relevant features. As EHR data continues to grow in complexity and volume, such systems may offer a scalable path forward for real-world deployment of personalized, AI-driven healthcare.

References

Emily Alsentzer, John R. Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew B. A. McDermott. Publicly available Clinical BERT embeddings. *arXiv:1904.03323 [cs.CL]*, 2019.

David Anderson, Michaela Anderson, Margret Bjarnadóttir, Stephen Mahar, and Shriyan Reyya. Paging dr. GPT: Extracting information from clinical notes to enhance patient predictions. *arXiv preprint arXiv:2504.12338*, 2025. URL https://arxiv.org/abs/2504.12338. 26 pages, cs.CL, cs.LG.

American Urological Association. PSA testing for the pretreatment staging and posttreatment management of prostate cancer - American Urological Association, 2013. URL https://www.auanet.org/guidelines/guidelines/prostate-specific-antigen-(psa)-best-practice-statement.

Mohsen Bayati, Mark Braverman, Michael Gillam, Karen M. Mack, George Ruiz, Mark S. Smith, and Eric Horvitz. Data-driven decisions for reducing readmissions for heart failure: general methodology and case study. *PLoS ONE*, 9(10):e109264, 2014. doi:10.1371/journal.pone.0109264.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: the long-document transformer. *arXiv:2004.05150 [cs.CL]*, 2020.

Mark K. Buyyounouski, Peter L. Choyke, Jesse K. McKenney, Oliver Sartor, Howard M. Sandler, Mahul B. Amin, Michael W. Kattan, and Daniel W. Lin. Prostate cancer - major changes in the American Joint Committee on Cancer eighth edition cancer staging manual. *CA: A Cancer Journal for Clinicians*, 67(3):245–253, 2017. doi:10.3322/caac.21391.

Jonathan H. Chen and Steven M. Asch. Machine learning and prediction in medicine — beyond the peak of inflated expectations. *New England Journal of Medicine*, 376(26):2507–2509, 2017. doi:10.1056/NEJMp1702071.

Sylvain DeLisle, Brett South, Jill A. Anthony, Ericka Kalp, Adi Gundlapallli, Frank C. Curriero, Greg E. Glass, Matthew Samore, and Trish M. Perl. Combining free text and structured Electronic Medical Record entries to detect acute respiratory infections. *PLoS ONE*, 5(10):e13377, 2010. doi:10.1371/journal.pone.0013377.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, 2019. Association for Computational Linguistics. doi:10.18653/v1/N19-1423.

Anjani Dhrangadhariya, Sebastian Otálora, Manfredo Atzori, and Henning Müller. Classification of noisy free-text prostate cancer pathology reports using natural language processing. In *Pattern Recognition. ICPR International Workshops and Challenges*, Lecture Notes in Computer Science, pages 154–166, Cham, 2021. Springer International Publishing. ISBN 978-3-030-68763-2. doi:10.1007/978-3-030-68763-2_12.

Author et al. Clinicalmamba: A large language model trained on longitudinal ehrs for contextualized clinical feature extraction. *npj Digital Medicine*, 7:150, 2024.

Author et al. Evaluation of deepseek r1 for complex medical scenario understanding in structured extraction tasks. *International Journal of Medical Informatics*, 2025a.

Author et al. Evaluating chatgpt, gemini and other large language models for structured extraction on synthetic clinical notes. *Journal of Artificial Intelligence in Clinical Medicine*, 2025b.

Dai Feng, Giuliana Cortese, and Richard Baumgartner. A comparison of confidence/credible interval methods for the area under the ROC curve for continuous diagnostic tests with small sample size. *Statistical Methods in Medical Research*, 26(6):2603–2621, 2017. doi:10.1177/0962280215602040.

Elizabeth Ford, John A Carroll, Helen E Smith, Donia Scott, and Jackie A Cassell. Extracting information from the text of electronic medical records to improve case detection: a systematic review. *Journal of the American Medical Informatics Association : JAMIA*, 23(5):1007–1015, 2016. doi:10.1093/jamia/ocv180.

Jason A. Fries, Ethan Steinberg, Saelig Khattar, Scott L. Fleming, Jose Posada, Alison Callahan, and Nigam H. Shah. Ontology-driven weak supervision for clinical entity classification in Electronic Health Records. *Nature Communications*, 12(1):2017, 2021. doi:10.1038/s41467-021-22328-4.

Sebastian Gehrmann, Franck Dernoncourt, Yeran Li, Eric T. Carlson, Joy T. Wu, Jonathan Welt, John Foote Jr, Edward T. Moseley, David W. Grant, Patrick D. Tyler, and Leo A. Celi. Comparing deep learning and concept extraction based methods for patient phenotyping from clinical narratives. *PLOS ONE*, 13(2):e0192360, 2018. doi:10.1371/journal.pone.0192360.

Marzyeh Ghassemi and Elaine Okanyene Nsoesie. In medicine, how do we machine learn anything real? *Patterns*, 3(1):100392, 2022. doi:10.1016/j.patter.2021.100392.

Milena A. Gianfrancesco, Suzanne Tamang, Jinoos Yazdany, and Gabriela Schmajuk. Potential biases in machine learning algorithms using Electronic Health Record data. *JAMA Internal Medicine*, 178 (11):1544–1547, 2018. doi:10.1001/jamainternmed.2018.3763.

William P. Glasheen, Tristan Cordier, Rajiv Gumpina, Gil Haugh, Jared Davis, and Andrew Renda. Charlson Comorbidity Index: ICD-9 update and ICD-10 translation. *American Health & Drug Benefits*, 12(4):188–197, 2019. ISSN 1942-2962.

Udo Hahn and Michel Oleynik. Medical information extraction in the age of deep learning. *Yearbook of Medical Informatics*, 29(1):208–220, 2020. doi:10.1055/s-0040-1702001.

Katharine E. Henry, David N. Hager, Peter J. Pronovost, and Suchi Saria. A targeted real-time early warning score (TREWScore) for septic shock. *Science Translational Medicine*, 7(299), 2015. doi:10.1126/scitranslmed.aab3719.

Chao-Chun Hsu, Shantanu Karnwal, Sendhil Mullainathan, Ziad Obermeyer, and Chenhao Tan. Characterizing the value of information in medical notes. *arXiv:2010.03574 [cs.CL]*, 2020.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021. URL `https://arxiv.org/abs/2106.09685`.

Jingwei Huang, Donghan M. Yang, Ruichen Rong, Kuroush Nezafati, Colin Treager, Zhikai Chi, Shidan Wang, Xian Cheng, Yujia Guo, Laura J. Klesse, Guanghua Xiao, Eric D. Peterson, Xiaowei Zhan, and Yang Xie. A critical assessment of using chatgpt for extracting structured data from clinical notes. *npj Digital Medicine*, 7:106, 2024. doi:10.1038/s41746-024-01079-8. URL `https://doi.org/10.1038/s41746-024-01079-8`.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. ClinicalBERT: modeling clinical notes and predicting hospital readmission. *arXiv:1904.05342 [cs.CL]*, 2020.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023. URL `https://arxiv.org/abs/2310.06825`.

Matthieu Komorowski, Leo A. Celi, Omar Badawi, Anthony C. Gordon, and A. Aldo Faisal. The Artificial Intelligence Clinician learns optimal treatment strategies for sepsis in intensive care. *Nature Medicine*, 24(11):1716–1720, 2018. doi:10.1038/s41591-018-0213-5.

Jingshu Liu, Zachariah Zhang, and Narges Razavian. Deep EHR: chronic disease prediction using medical notes. In *Proceedings of the 3rd Machine Learning for Healthcare Conference*, pages 440–464. PMLR, 2018. URL `https://proceedings.mlr.press/v85/liu18b.html`. ISSN: 2640-3498.

Denis McInerney, Geoffrey Young, Jan-Willem van de Meent, and Byron Wallace. CHiLL: Zero-shot custom interpretable feature extraction from clinical notes with large language models. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8477–8494, Singapore, December 2023. Association for Computational Linguistics. doi:10.18653/v1/2023.findings-emnlp.568. URL `https://aclanthology.org/2023.findings-emnlp.568/`.

Robert Moskovitch, Fernanda Polubriaginof, Aviram Weiss, Patrick Ryan, and Nicholas Tatonetti. Procedure prediction from symbolic Electronic Health Records via time intervals analytics. *Journal of Biomedical Informatics*, 75:70–82, 2017. doi:10.1016/j.jbi.2017.07.018.

Robert G. Newcombe. Confidence intervals for an effect size measure based on the Mann–Whitney statistic. Part 2: asymptotic methods and evaluation. *Statistics in Medicine*, 25(4):559–573, 2006. doi:10.1002/sim.2324.

Thuy Nguyen, Ying Geng, Elizabeth Proctor, Valerie Arboleda, James E Gill, Matthew Snyder, Julianne D Brooks, Meenakshi Kaur, Corey W Arnold, and Hector Wilhalme. Generating high-quality data abstractions from scanned clinical pathology reports. *BMJ Open*, 10(11):e037740, 2020.

Ziad Obermeyer and Ezekiel J. Emanuel. Predicting the future — big data, machine learning, and clinical medicine. *The New England Journal of Medicine*, 375(13):1216–1219, 2016. doi:10.1056/NEJMp1606181.

Mack Roach, Gerald Hanks, Howard Thames, Paul Schellhammer, William U. Shipley, Gerald H. Sokol, and Howard Sandler. Defining biochemical failure following radiotherapy with or without hormonal therapy in men with clinically localized prostate cancer: recommendations of the RTOG-ASTRO Phoenix Consensus Conference. *International Journal of Radiation Oncology, Biology, Physics*, 65(4): 965–974, 2006. doi:10.1016/j.ijrobp.2006.04.029.

Kenneth D. Roe, Vibhu Jawa, Xiaohan Zhang, Christopher G. Chute, Jeremy A. Epstein, Jordan Matelsky, Ilya Shpitser, and Casey Overby Taylor. Feature engineering with clinical expert knowledge: a case study assessment of machine learning model complexity and performance. *PLOS ONE*, 15(4): e0231300, 2020. doi:10.1371/journal.pone.0231300.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv:1910.01108 [cs.CL]*, 2020.

William A. See. Postoperative nomogram predicting risk of recurrence after radical cystectomy for bladder cancer. *Urologic Oncology: Seminars and Original Investigations*, 25(3):275, 2007. doi:10.1016/j.urolonc.2007.03.011.

Laleh Seyyed-Kalantari, Haoran Zhang, Matthew B. A. McDermott, Irene Y. Chen, and Marzyeh Ghassemi. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in underserved patient populations. *Nature Medicine*, 27(12):2176–2182, 2021. doi:10.1038/s41591-021-01595-0.

Yuqi Si, Jingqi Wang, Hua Xu, and Kirk Roberts. Enhancing clinical concept extraction with contextual embeddings. *Journal of the American Medical Informatics Association*, 26(11):1297–1304, 2019. doi:10.1093/jamia/ocz096.

Anima Singh, Girish Nadkarni, Omri Gottesman, Stephen B. Ellis, Erwin P. Bottinger, and John V. Guttag. Incorporating temporal EHR data in predictive models for risk stratification of renal function deterioration. *Journal of Biomedical Informatics*, 53:220–228, 2015. doi:10.1016/j.jbi.2014.11.005.

Andrew J. Stephenson, Michael W. Kattan, James A. Eastham, Zohar A. Dotan, Fernando J. Bianco, Hans Lilja, and Peter T. Scardino. Defining biochemical recurrence of prostate cancer after radical prostatectomy: a proposal for a standardized definition. *Journal of Clinical Oncology*, 24(24):3973–3978, 2006. doi:10.1200/JCO.2005.04.0756.

Marion Verduijn, Lucia Sacchi, Niels Peek, Riccardo Bellazzi, Evert de Jonge, and Bas A.J.M. de Mol. Temporal abstraction for feature extraction: a comparative case study in prediction from intensive care monitoring data. *Artificial Intelligence in Medicine*, 41(1):1–12, 2007. doi:10.1016/j.artmed.2007.06.003.

Qingyu Wei, Yifan Peng, Xin Wang, and Lucila Ohno-Machado Wang. Gatortron: A large clinical language model for specialized ehr extraction. *Journal of Biomedical Informatics*, 117:103759, 2021.

Jenna Wiens, John Guttag, and Eric Horvitz. Patient risk stratification with time-varying parameters: A multitask learning approach. *Journal of Machine Learning Research*, 17(79):1–23, 2016. URL `http://jmlr.org/papers/v17/15-177.html`.

Yan Xu, Kai Hong, Junichi Tsujii, and Eric I-Chao Chang. Feature engineering combined with machine learning and rule-based methods for structured information extraction from narrative clinical discharge summaries. *Journal of the American Medical Informatics Association*, 19(5):824–832, 2012. doi:10.1136/amiajnl-2011-000776.

Di Zhao and Chunhua Weng. Combining PubMed knowledge and EHR data to develop a weighted bayesian network for pancreatic cancer prediction. *Journal of Biomedical Informatics*, 44(5):859–868, 2011. doi:10.1016/j.jbi.2011.05.004.

Alice Zheng and Amanda Casari. *Feature Engineering for Machine Learning*. O'Reilly Media, Inc., 2018. ISBN 1-4919-5324-1. URL https://www.oreilly.com/library/view/feature-engineering-for/9781491953235/. ISBN: 9781491953242.

Appendix

A SNOW Feature Table

Table 3: Feature Specifications for Prostate Cancer Prediction Model Generated by SNOW

Feature Name	Specific Subgroups	Description	Instructions	Agg.	Agg. Source
gleason score primary	All 14 prostate regions*	Primary Gleason pattern (1-5) for each specific prostate region. The primary pattern represents the predominant histological pattern observed in the tumor tissue. Higher primary patterns indicate more aggressive disease.	Extract the first number in the Gleason score pattern (e.g., '4' from 'Gleason 4+3=7') for each specific region. If no cancer is found in a region (e.g., 'BENIGN PROSTATIC GLANDS AND STROMA'), code as 0.	FALSE	–
gleason score secondary	All 14 prostate regions*	Secondary Gleason pattern (1-5) for each specific prostate region. The secondary pattern represents the second most common histological pattern in the tumor tissue. Higher secondary patterns indicate more aggressive disease.	Extract the second number in the Gleason score pattern (e.g., '3' from 'Gleason 4+3=7') for each specific region. If no cancer is found in a region (e.g., 'BENIGN PROSTATIC GLANDS AND STROMA'), code as 0.	FALSE	–
gleason score sum	All 14 prostate regions*	Total Gleason score (2-10) for each specific prostate region. The sum of primary and secondary Gleason patterns, with higher scores indicating more aggressive disease.	Extract the total Gleason score (e.g., '7' from 'Gleason 4+3=7') for each specific region. If no cancer is found in a region (e.g., 'BENIGN PROSTATIC GLANDS AND STROMA'), code as 0.	FALSE	–

Continued on next page

Table 3 – *Continued from previous page*

Feature Name	**Specific Subgroups**	**Description**	**Instructions**	**Agg.**	**Agg. Source**
tumor percentage	All 14 prostate regions*	Percentage of core involved by tumor for each specific prostate region. Higher percentages indicate greater tumor burden and potentially more aggressive disease.	Extract the percentage value from phrases like 'COMPRISING 50% OF THE CORE' or 'INVOLVING 75% OF THE CORE'. If no cancer is found in a region (e.g., 'BENIGN PROSTATIC GLANDS AND STROMA'), code as 0.	FALSE	–
cancer presence	All 14 prostate regions*	Binary indicator of cancer presence in each specific prostate region.	Code as 1 if cancer is found in a region (e.g., 'PROSTATIC ADENOCARCINOMA'), code as 0 if no cancer is found (e.g., 'BENIGN PROSTATIC GLANDS AND STROMA').	FALSE	–
total cores count	–	Total number of biopsy cores sampled. Indicates the extent of sampling.	Count the total number of regions sampled in the report.	FALSE	–
intraductal carcinoma presence	–	Presence of intraductal carcinoma, which is associated with more aggressive disease and poorer outcomes.	Code as 1 if the phrase 'INTRADUCTAL CARCINOMA' appears in any region description, otherwise code as 0.	FALSE	–
prostate volume	–	Volume of the prostate in cubic centimeters as measured during biopsy. Larger prostates may have different disease characteristics.	Extract the numerical value following 'Volume =' or similar phrases in the operative findings section.	FALSE	–
psa pre biopsy	–	PSA level immediately before biopsy. Higher levels may indicate more advanced disease.	Extract the most recent PSA value mentioned in the clinical history section before the biopsy date.	FALSE	–

Continued on next page

Table 3 – *Continued from previous page*

Feature Name	**Specific Subgroups**	**Description**	**Instructions**	**Agg.**	**Agg. Source**
max gleason score primary	–	Maximum primary Gleason pattern across all sampled regions. Represents the highest grade of cancer found in any region of the prostate.	Calculate the maximum value of primary Gleason pattern across all sampled regions.	TRUE	gleason score primary (all regions)
max gleason score secondary	–	Maximum secondary Gleason pattern across all sampled regions. Represents the highest secondary grade of cancer found in any region of the prostate.	Calculate the maximum value of secondary Gleason pattern across all sampled regions.	TRUE	gleason score secondary (all regions)
max gleason score sum	–	Maximum total Gleason score across all sampled regions. Represents the highest overall grade of cancer found in any region of the prostate.	Calculate the maximum value of total Gleason score across all sampled regions.	TRUE	gleason score sum (all regions)
max tumor percentage	–	Maximum percentage of core involved by tumor across all sampled regions. Indicates the highest tumor burden in any single core.	Calculate the maximum percentage of core involved by tumor across all sampled regions.	TRUE	tumor percentage (all regions)
positive cores count	–	Number of biopsy cores positive for cancer. Higher numbers indicate more widespread disease.	Count the number of regions with any cancer finding (i.e., not 'BENIGN PROSTATIC GLANDS AND STROMA').	TRUE	cancer presence (all regions)
percentage positive cores	–	Percentage of sampled cores positive for cancer. Higher percentages indicate more widespread disease.	Calculate (positive cores count / total cores count) * 100.	TRUE	positive cores count, total cores count

Continued on next page

Table 3 – *Continued from previous page*

Feature Name	**Specific Subgroups**	**Description**	**Instructions**	**Agg.**	**Agg. Source**
bilateral disease	–	Presence of cancer in both left and right sides of the prostate, indicating more widespread disease.	Code as 1 if cancer is found in at least one region on both the left and right sides, otherwise code as 0.	TRUE	cancer presence (L/R regions)

*14 prostate regions: left/right × (apex medial, apex lateral, mid medial, mid lateral, base medial, base lateral, anterior apex)