Predicting Glucose Test Ordering in Hospitalized Patients Using Temporal Models of Clinical Context Embeddings

Joud El-Shawa[1,2], **Elham Bagheri**[1,2], **Amol Verma**[3], **Yalda Mohsenzadeh**[1,2*]

[1] Vector Institute for Artificial Intelligence, Toronto, ON M5G 0C6, Canada
[2] Department of Computer Science, Western University, London, ON N6A 3K7, Canada
[3] Unity Health Toronto, Toronto, ON M5G 2C4, Canada
[*]ymohsenz@uwo.ca

Abstract

The overuse of laboratory tests is a persistent challenge in healthcare, driving unnecessary costs, patient discomfort, and low-value care. Glucose testing, one of the most common diagnostics, exemplifies this issue in hospital settings. We present a deep learning framework that integrates structured and unstructured electronic medical record data to predict whether a glucose test will be ordered in the next AM/PM time bin. Using multi-hospital data from the GEMINI dataset, we combine Long Short-Term Memory models with Clinical BioBERT embeddings to capture both the timing and clinical context of testing. On held-out test data, our best model achieved ROC-AUC of 0.92 and PR-AUC of 0.67, and generalized across sites in leave-one-hospital-out evaluation (ROC-AUC 0.84). Embedding-based models outperformed traditional feature representations, though adding more tests and vitals did not always yield further gains. By contrast, introducing a simple temporal recency cue (bin counter) improved performance. An exploratory regression task for predicting glucose values performed worse, likely due to class imbalance and reliance on forward-filled values; Random Forest achieved R^2 of 0.80 under masked evaluation, indicating a need for more frequent or diverse test data. Predicting laboratory test ordering is the first step toward evaluating the usefulness of laboratory test use and establishes a foundation for future real-time decision support to reduce unnecessary lab use in hospitals.

Introduction

Diagnostic testing is central to clinical decision-making, yet overuse remains a persistent problem in hospitals, leading to unnecessary costs, patient harm, and low-value care (Müskens et al. 2022; Vrijsen et al. 2020). Estimates suggest that roughly 30% of laboratory tests are unnecessary, contributing to false positives, additional procedures, and reduced patient satisfaction. Initiatives such as Choosing Wisely Canada have issued hundreds of recommendations to reduce overuse, but the Canadian Institute for Health Information (CIHI) notes that system-level changes are still needed to achieve lasting improvements (2022). Glucose testing, one of the most frequently ordered labs, exemplifies this challenge and is the focus of this study.

Traditional statistical methods often fail to capture the complex temporal and contextual relationships inherent in electronic medical record (EMR) data; recent advances in deep learning offer promising solutions, particularly through architectures that can capture temporal dependencies and process complex clinical data. Recurrent Neural Networks (RNNs), such as Long Short-Term Memory (LSTM) networks, excel at modelling sequential data and capturing long-term dependencies, making them well suited for predicting temporal patterns in healthcare, such as glucose testing (Rajkomar et al. 2018; Harutyunyan et al. 2019).

In parallel, advancements in Natural Language Processing (NLP) have enabled the extraction of structured insights from unstructured clinical text. Foundational models, such as Clinical BioBERT, have demonstrated the ability to extract meaningful embeddings from unstructured medical data, offering further potential for improving predictive accuracy (Alsentzer et al. 2019). In clinical settings, where data is often missing, inconsistently formatted, or sparsely documented, these transformer-based embeddings provide a standardized method of encoding heterogeneous information.

This research uses the GEMINI dataset, a large-scale multi-hospital database encompassing over 2.4 million admissions and 12 billion data points across over 30 hospitals (Verma et al. 2021; GEMINI 2025). GEMINI provides detailed information, including laboratory tests, vital signs, imaging reports, and various clinical variables, making it a valuable resource for predictive modelling. By integrating data from this dataset, we aim to develop a machine learning framework capable of accurately predicting whether a glucose test will be *ordered* in the next AM or PM (twice-daily) time bin for hospitalized patients. Specifically, we explore the use of Clinical BioBERT embeddings and LSTM networks to model temporal patterns and contextual relationships in EMR data.

This study offers two key contributions. First, it introduces a preprocessing pipeline designed to address issues such as missing data and class imbalance within electronic health record (EHR) datasets. Second, it offers a detailed analysis of how different data modalities contribute to the prediction of lab test utilization, offering insights into how diverse EHR data can support predictive modelling in healthcare more broadly. By systematically exploring di-

verse inputs and applying advanced methodologies within this specific task, this research underscores both the opportunities and limitations of artificial intelligence in predictive healthcare.

Background and Related Work

Machine learning (ML) models have shown promise in healthcare prediction tasks, including laboratory testing. Zale et al. (2022) predicted in-hospital blood glucose levels using demographic, lab, and vital sign data, with Random Forest outperforming methods based only on the latest glucose value. Cheng, Prasad, and Engelhardt (2019) applied reinforcement learning to optimize lab ordering policies, reducing tests by 44% in MIMIC-III (Johnson et al. 2016) while balancing clinical value and cost. Islam et al. (2021) developed a deep neural network to recommend relevant tests from EHR data, addressing over- and underuse, while Yu et al. (2020) predicted lab results to reduce unnecessary testing by over 20% without compromising accuracy. Using the GEMINI dataset (Verma et al. 2021; GEMINI 2025), Weinerman et al. (2023) identified 21 frequently overused tests for targeted reduction, and Fralick et al. (2021) predicted hypoglycemia risk with ML and NLP on clinical notes.

However, prior work often suffers from limited handling of missing data (Zale et al. 2022), lack of temporal modelling (Islam et al. 2021), or single-institution datasets (Yu et al. 2020), restricting generalizability. Unlike studies that primarily predict future laboratory values, we predict the near-term *ordering* of a glucose test, directly targeting utilization. We address these gaps by (1) applying preprocessing methods that are robust to missing values, (2) incorporating temporal sequences, and (3) using GEMINI's multihospital data for external validity. This situates our work alongside emerging literature on foundation EHR models, which similarly leverage embeddings to improve robustness and generalization across healthcare tasks (Guo et al. 2024).

Specifically, we (i) adopt the utilization framing of Cheng, Prasad, and Engelhardt (2019) and select operating points that prioritize recall at clinically acceptable specificity; (ii) extend Fralick et al. (2021) by encoding clinical context with transformer embeddings (Clinical BioBERT) instead of hand-engineered text features; and (iii) connect to overuse reduction goals from Weinerman et al. (2023) by grounding the task in GEMINI-mapped tests and evaluating generalizability across hospitals.

Methods

We developed an ML system to predict the occurrence of glucose laboratory tests, informed by preliminary data analysis and expert consultation. All analysis was performed in Python (Van Rossum and Drake 2009) within JupyterLab (Kluyver et al. 2016), using SQLAlchemy (Bayer 2012) for data retrieval. Our experiments compare four pipelines (A–D) differing in the inclusion and representation of multimodal data (structured, embedded, extended embedded).

Data Preprocessing

Labs. We extracted all glucose-related tests from the GEMINI dataset (Verma et al. 2021; GEMINI 2025) using mapped codes from the clinical team. Implausible values above the 99th percentile (by z-score) were removed. Each patient visit was split into twice-daily bins (AM/PM) based on observed collection patterns and expert input. A binary *performed* label indicated test occurrence; the prediction target was a shifted version of this label, indicating whether a test would occur in the next bin. All bins within a patient's length of stay were generated, ensuring temporal continuity; missing bins were inserted and marked with *performed* = 0. Multiple glucose results in the same bin were averaged. Missing result values were filled differently depending on their position: initial missing values were replaced with a normal reference value of 4.5 mmol/L, and intermediate missing values were forward-filled from the last observed measurement.

Vitals. Vital signs were mapped by code, filtered to match the Labs cohort, and binned identically. For Pipelines A and B, we used only numeric vitals (e.g., blood pressure, temperature, pulse). Non-numeric vitals were deferred to embedding-based pipelines. A binary *performed* label for each vital indicated whether a measurement occurred in a bin. Missing bins were explicitly inserted; intermediate gaps were forward-filled from the last observed value, while initial missing values were left unfilled due to variability across vitals. In bins without a recorded measurement, the original value column was set to -1 so the model could distinguish true values from missing data. Multiple measurements within a bin were averaged.

Admissions. The admissions table contained demographic and visit-level information. For Pipelines A and B, we included a subset of categorical and numeric fields that require minimal manual preprocessing (e.g., gender, age, admit category), standardizing categories before one-hot encoding. More complex free-text and high-cardinality features were reserved for embedding pipelines.

Merging. Labs and Vitals were outer-joined on patient ID and bin, with missing values handled according to the preprocessing rules. Admissions data were left-joined onto the merged dataframe. Outlier stays (bin count) above the 99th percentile were excluded. Numeric features were minmax scaled to [0,1], while glucose measurements robust scaled (Pedregosa et al. 2011) to reduce the influence of extreme values, which occur more frequently in glucose tests.

All scalers (and, where applicable, PCA) were fit on the training split only and then applied to validation/test splits to prevent train–test contamination. All analyses used deidentified data from the GEMINI dataset under research ethics board approval, with secure governance and access controls in place (Verma et al. 2021; GEMINI 2025).

Embeddings Preprocessing. Pipelines B–D used embeddings to encode structured data into text paragraphs, which were embedded using Clinical BioBERT (Alsentzer et al. 2019). Principal Component Analysis (PCA) (Jolliffe and Cadima 2016) was applied to embeddings to retain 99% of

Pipe	Labs	Vitals	Adm	BC
A	Gluc/Cls	Map/Cls	Some/Cls	No
B	Gluc/Emb	Map/Emb	Some/Emb	No
C	All L/Emb	All V/Emb	All A/Emb	No
D	All L/Emb	All V/Emb	All A/Emb	Yes

Table 1: Setup of different pipeline configurations.

Abbreviations: Adm = admissions; BC = bin counter, Gluc = glucose-only; All L = all labs; All V = all vitals; All A = all admissions; Map = mapped numerics; Cls = classic tabular; Emb = Clinical BioBERT embeddings.

P	Acc	Prec	Rec	Spec	F1	ROC	PR
A	0.85	0.44	0.77	0.87	0.56	0.87	0.60
B	0.87	0.48	0.74	0.89	0.59	0.89	0.60
C	0.83	0.37	0.75	0.84	0.50	0.87	0.57
D	**0.88**	**0.48**	**0.81**	**0.89**	**0.60**	**0.92**	**0.67**

Table 2: Performance Metrics on Test Data for Pipelines A, B, C, and D

Abbreviations: P = Pipeline, Acc = Accuracy, Prec = Precision, Rec = Recall, Spec = Specificity, ROC = Receiver-Operating Characteristic Curve, PR = Precision–Recall Area Under the Curve (AUC).

the variance while reducing dimensionality for downstream modelling and mitigating overfitting.

Pipelines C and D incorporated additional sources: all lab tests specified by our subject-matter expert, unmapped vitals, and additional admissions features. Values were aggregated with counts, averages, minima, maxima, and reference ranges for numeric variables, and categorical distributions for non-numeric ones. To avoid token limits, Labs, Vitals, and Admissions paragraphs were embedded separately, PCA-reduced, and concatenated. Pipeline D additionally included the bin counter as a Robust-scaled numeric input.

Sequence Formation. Sequences of two bins were created to target near-term decision support and align with common hospital practice. Visits with fewer than three bins were excluded. Splitting was done at the *patient_id* level to prevent leakage, with 65%/19%/16% train/test/validation splits fixed across pipelines. The final dataset contained approximately 84,000 unique patients, 133,000 hospital visits, and 2.2 million generated sequences.

Model Training. Class imbalance (approximately 6:1 ratio of not-performed to performed tests) was addressed by computing class weights and passing them to the model during training to penalize misclassification of the minority class. Each pipeline (Table 1) was trained using an LSTM (Hochreiter and Schmidhuber 1997) implemented in Keras (Chollet 2024) with TensorFlow (Abadi et al. 2016), binary cross-entropy loss, Adam optimizer (Kingma and Ba 2014), and sigmoid output activation. We performed grid search, and the settings yielding the highest validation performance were selected.

Pipelines C and D were additionally tuned via Bayesian optimization (Snoek, Larochelle, and Adams 2012) to explore more complex architectures. While this approach efficiently sampled broader configurations, performance gains were marginal ($< 1\%$) and often came at the cost of greater complexity and longer training. Given the need for interpretability and deployment efficiency, final models used the simpler manually-tuned configurations.

Model Architecture and Hyperparameter Search. To evaluate generalizability, Pipeline A underwent Leave-One-Hospital-Out cross-validation (LOHO CV) using GEMINI's (Verma et al. 2021; GEMINI 2025) multi-hospital dataset. Because patients could appear in multiple hospitals, each patient was assigned to the hospital with the most bins, ensuring no cross-hospital overlap. For each LOHO fold, data from one hospital served as the test set while remaining hospitals formed the training/validation sets, with preprocessing and sequence formation repeated per fold.

Results

Data Analysis. Exploratory analysis of GEMINI revealed distinct temporal and hospital-level patterns in glucose testing and vital measurements. Glucose testing peaks occurred between 5–8 AM, while vital signs were measured most frequently at 8 AM and 8 PM. Testing frequency generally increased over time, with a notable dip during 2020–2021, likely due to COVID-19. Two hospitals (102, 115) were excluded for insufficient data.

Prediction Performance. Table 2 shows the best test-set results for each pipeline (A–D), reported at the highest validation ROC-AUC, using a 0.5 threshold on model outputs.

External Validation Across Hospitals. Pipeline A was externally validated using Leave-One-Hospital-Out CV. Embedding-based validation was not feasible due to high-performance computing (HPC) resource limitations, including GPU memory and runtime constraints. Nevertheless, the model achieved an accuracy of **0.85** $\pm$ **0.08** (95% CI: 0.79–0.92), precision of **0.42** $\pm$ **0.16** (95% CI: 0.29–0.55), recall of **0.72** $\pm$ **0.07** (95% CI: 0.66–0.78), specificity of **0.87** $\pm$ **0.09** (95% CI: 0.80–0.94), F1-score of **0.51** $\pm$ **0.15** (95% CI: 0.38–0.63), PR-AUC of **0.51** $\pm$ **0.08** (95% CI: 0.45–0.57), and ROC-AUC of **0.84** $\pm$ **0.05** (95% CI: 0.80–0.88), indicating good generalization to unseen hospitals.

Exploring Regression

In addition to classification, we explored predicting actual glucose values to provide a richer clinical context. This was framed as a regression task using Pipeline A features for interpretability. A Random Forest (RF) (Breiman 2001) regressor underwent grid search, yielding the best configuration of *n_est* = 500, *max_depth* = 20, *min_split* = 10, *min_leaf* = 4. On the test set (with masked evaluation applied to bins with true observed values to avoid bias from forward-filled glucose), the model achieved a mean absolute error of 0.58, a mean squared error of 1.59, and a coefficient of determination of 0.80, with a runtime of 1,682 seconds.

Initial LSTM regression attempts were unstable, potentially due to short sequences, high variability, and forward-

filling effects, and are not reported here. The RF model outperformed LSTM in this setting, possibly because it evaluates samples independently rather than relying on potentially misleading short-term trends.

While regression performance was modest, likely due to the inherent imbalance in glucose measurements, further work should explore more balanced, frequently measured tests (e.g., sodium) for comparison. Extending the pipeline to jointly output binary predictions and continuous values could enhance clinical decision-making, including downstream classification into normal/abnormal ranges

Discussion

Our results indicate a clear improvement in performance when transitioning from Pipeline A (which uses classic mapped glucose data, mapped vitals, and select admissions features as input) to Pipeline B (where the same data is represented in paragraph form and embedded using NLP techniques). Representing structured clinical data as paragraphs and applying embeddings not only boosts performance, but also captures patient history more accurately, without relying on assumptions such as normal or carried-forward values for missing measurements. In contrast, traditional preprocessing approaches, like imputing "normal" results or forward-filling previous values, can introduce inconsistencies and fail to reflect a patient's evolving clinical state; embeddings instead leverage all available information at each time point, even when partial or incomplete, enabling a more realistic understanding of the patient's condition.

However, despite integrating additional tests, unmapped vitals, and expanded admissions data, Pipeline C did not yield further improvements. This could be due to the increased complexity of the embeddings, which could have introduced noise or irrelevant patterns that the simple LSTM architecture struggled to effectively model and interpret. The model may have been unable to fully leverage the added information, suggesting that a more complex model architecture or additional regularization techniques could be necessary to better capture the nuances of the richer embeddings.

Finally, given the limitation of short patient stays and the restriction on the sequence length by the GEMINI team, the inclusion of the bin counter variable in Pipeline D proves to be beneficial. This variable provides additional context about patient history, contributing to more accurate predictions. The performance in Pipeline D suggests that structuring temporal information explicitly, rather than relying solely on embeddings and implied relationships through sequences, enhances predictive accuracy.

Given the clinical priority of avoiding missed necessary tests, evaluation emphasized sensitivity (recall) over other metrics, aligning with subject-matter expert guidance. Short sequences (2 bins) matched clinical workflows where rapid decision support is essential, and the bin counter partially offsets limited history. Future work could explore ensembles or transformer-based (Vaswani et al. 2017) models to flexibly handle variable-length sequences.

Overall, while each pipeline was carefully designed to enhance predictive performance, our results emphasize the need to balance data complexity and model capacity. Increasing the complexity of embeddings does not guarantee better performance, as ensuring that the model can effectively extract relevant information is equally important.

This study has several limitations. First, hospital-level differences in lab ordering practices, calibration, and reference ranges present potential threats to validity. Second, although using GEMINI-mapped tests in Pipeline A helped reduce some inconsistencies, data quality issues such as outliers and missing units remain. Third, while embeddings helped mitigate the impact of missing or inconsistent values by leveraging all available information, the model's availability to generalize to datasets with different data structures has not been evaluated. In the future, enhanced preprocessing by the GEMINI medical team, such as refining data collection processes or addressing current data quality issues, could further enhance model utility.

Our approach, which is the first to predict lab test occurrence in the GEMINI dataset, combines LSTMs with NLP embeddings for structured hospital data, offering potential to reduce unnecessary testing, cut costs, and improve care. External validation showed strong results across hospitals, though its applicability to hospitals with different patient demographics may require further refinement and validation.

Conclusion

The over-utilization of laboratory tests remains an ongoing challenge in healthcare, driving costs, hospital-acquired anemia, heightened patient anxiety, and low-value care (Müskens et al. 2022; Shaik et al. 2024). While initiatives requiring manual labour have been implemented in Canada to address this issue, system-level changes are crucial for achieving broader and more sustainable improvements (Canadian Institute for Health Information 2022).

In this paper, we presented a novel framework for predicting laboratory test occurrences using the GEMINI dataset, combining LSTM networks with NLP-based embeddings to integrate structured and unstructured clinical data. This approach addressed common challenges such as missing values and class imbalance, improving predictive accuracy while capturing richer temporal and contextual information. Beyond classification, we explored regression for predicting test values, noting the impact of data imbalance. Future work will extend this work to more frequent tests like sodium, which offer more balanced datasets and could help determine whether the lower glucose performance was due to imbalance or inherent task complexity.

Additional directions include ensemble models for variable sequence lengths, which can enable dynamic model selection based on the historical data available for each patient to optimize predictive accuracy while maintaining clinical practicality. Integrating direct clinician feedback into the evaluation process will be essential to refine predictions and ensure alignment with real-world decision-making workflows, thus improving both trust and practical utility in clinical settings. These contributions provide a strong basis for AI applications in healthcare, offering novel approaches for modelling clinical data and setting the stage for further advancements in predictive healthcare analytics.

References

Abadi, M.; Agarwal, A.; Barham, P.; Brevdo, E.; Chen, Z.; Citro, C.; Corrado, G. S.; Davis, A.; Dean, J.; Devin, M.; et al. 2016. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*.

Alsentzer, E.; Murphy, J. R.; Boag, W.; Weng, W.-H.; Jin, D.; Naumann, T.; and McDermott, M. B. A. 2019. Publicly Available Clinical BERT Embeddings. *arXiv:1904.03323 [cs.CL]*.

Bayer, M. 2012. SQLAlchemy. In Brown, A.; and Wilson, G., eds., *The Architecture of Open Source Applications Volume II: Structure, Scale, and a Few More Fearless Hacks*. aosabook.org.

Breiman, L. 2001. Random Forests. *Machine Learning*, 45(1): 5–32.

Canadian Institute for Health Information. 2022. Overuse of tests and treatments in Canada.

Cheng, L.-F.; Prasad, N.; and Engelhardt, B. E. 2019. An optimal policy for patient laboratory tests in intensive care units. In *Pacific Symposium on Biocomputing 2019*, 320–331.

Chollet, F. 2024. Keras, within TensorFlow (Version 2.11.0) [Software component].

Fralick, M.; Dai, D.; Pou-Prom, C.; Verma, A. A.; and Mamdani, M. 2021. Using machine learning to predict severe hypoglycaemia in hospital. *Diabetes, Obesity and Metabolism*, 23(10): 2311–2319.

GEMINI. 2025. The GEMINI Database. Accessed: 2025-08-12.

Guo, L. L.; Fries, J.; Steinberg, E.; Fleming, S. L.; Morse, K.; Aftandilian, C.; Posada, J.; Shah, N.; and Sung, L. 2024. A multicenter study on the adaptability of a shared foundation model for electronic health records. *npj Digital Medicine*, 7(1): 171.

Harutyunyan, H.; Khachatrian, H.; Kale, D. C.; Steeg, G. V.; and Galstyan, A. 2019. Multitask learning and benchmarking with clinical time series data. *Scientific Data*, 6(1): 96.

Hochreiter, S.; and Schmidhuber, J. 1997. Long short-term memory. *Neural computation*, 9(8): 1735–1780.

Islam, M. M.; Poly, T. N.; Yang, H.-C.; and Li, Y.-C. J. 2021. Deep into laboratory: An artificial intelligence approach to recommend laboratory tests. *Diagnostics*, 11(6): 990.

Johnson, A. E.; Pollard, T. J.; Shen, L.; Lehman, L.-w. H.; Feng, M.; Ghassemi, M.; Moody, B.; Szolovits, P.; Anthony Celi, L.; and Mark, R. G. 2016. MIMIC-III, a freely accessible critical care database. *Scientific data*, 3(1): 1–9.

Jolliffe, I. T.; and Cadima, J. 2016. Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society*.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kluyver, T.; Ragan-Kelley, B.; Pérez, F.; Granger, B.; Bussonnier, M.; Frederic, J.; Kelley, K.; Hamrick, J.; Grout, J.; Corlay, S.; Ivanov, P.; Avila, D.; Abdalla, S.; and Willing, C. 2016. Jupyter Notebooks – a publishing format for reproducible computational workflows. In Loizides, F.; and Schmidt, B., eds., *Positioning and Power in Academic Publishing: Players, Agents and Agendas*, 87 – 90. IOS Press.

Müskens, J. L. J. M.; Kool, R. B.; van Dulmen, S. A.; and Westert, G. P. 2022. Overuse of diagnostic testing in healthcare: A systematic review. *BMJ Quality & Safety*, 31(1): 54–63.

Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; and Duchesnay, E. 2011. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.*, 12: 2825–2830.

Rajkomar, A.; Oren, E.; Chen, K.; Dai, A. M.; Hajaj, N.; Hardt, M.; Liu, P. J.; Liu, X.; Marcus, J.; Sun, M.; et al. 2018. Scalable and accurate deep learning with electronic health records. *NPJ digital medicine*, 1(1): 18.

Shaik, T.; Mahmood, R.; Kanagala, S. G.; Kaur, H.; Mendpara, V.; Gupta, V.; Aggarwal, P.; Anamika, F.; Garg, N.; and Jain, R. 2024. Lab testing overload: A comprehensive analysis of overutilization in hospital-based settings. *Proceedings (Baylor University Medical Center)*, 37(2): 312–316.

Snoek, J.; Larochelle, H.; and Adams, R. P. 2012. Practical Bayesian optimization of machine learning algorithms. In *Advances in neural information processing systems*, volume 25.

Van Rossum, G.; and Drake, F. L. 2009. *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace. ISBN 1441412697.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in neural information processing systems*, volume 30.

Verma, A. A.; Pasricha, S. V.; Jung, H. Y.; Kushnir, V.; Mak, D. Y.; Koppula, R.; Guo, Y.; Kwan, J. L.; Lapointe-Shaw, L.; Rawal, S.; et al. 2021. Assessing the quality of clinical and administrative data extracted from hospitals: the General Medicine Inpatient Initiative (GEMINI) experience. *Journal of the American Medical Informatics Association*, 28(3): 578–587.

Vrijsen, B. E. L.; Naaktgeboren, C. A.; Vos, L. M.; van Solinge, W. W.; Kaasjager, H. A. H.; and ten Berg, M. J. 2020. Inappropriate laboratory testing in internal medicine inpatients: Prevalence, causes and interventions. *Annals of Medicine and Surgery (London)*, 51: 48–53.

Weinerman, A. S.; Guo, Y.; Saha, S.; Yip, P. M.; Lapointe-Shaw, L.; Fralick, M.; Kwan, J. L.; MacMillan, T. E.; Liu, J.; Rawal, S.; Sheehan, K. A.; Simons, J.; Tang, T.; Bhatia, S.; Razak, F.; and Verma, A. A. 2023. Data-driven approach to identifying potential laboratory overuse in general internal medicine (GIM) inpatients. *BMJ Open Quality*, 12: Article e002261.

Yu, L.; Li, L.; Bernstam, E.; and Jiang, X. 2020. A deep learning solution to recommend laboratory reduction strategies in ICU. *International Journal of Medical Informatics*, 144: Article 104282.

Zale, A. D.; Abusamaan, M. S.; McGready, J.; and Mathioudakis, N. 2022. Development and validation of a machine learning model for classification of next glucose measurement in hospitalized patients. *eClinicalMedicine*, 44: Article 101290.