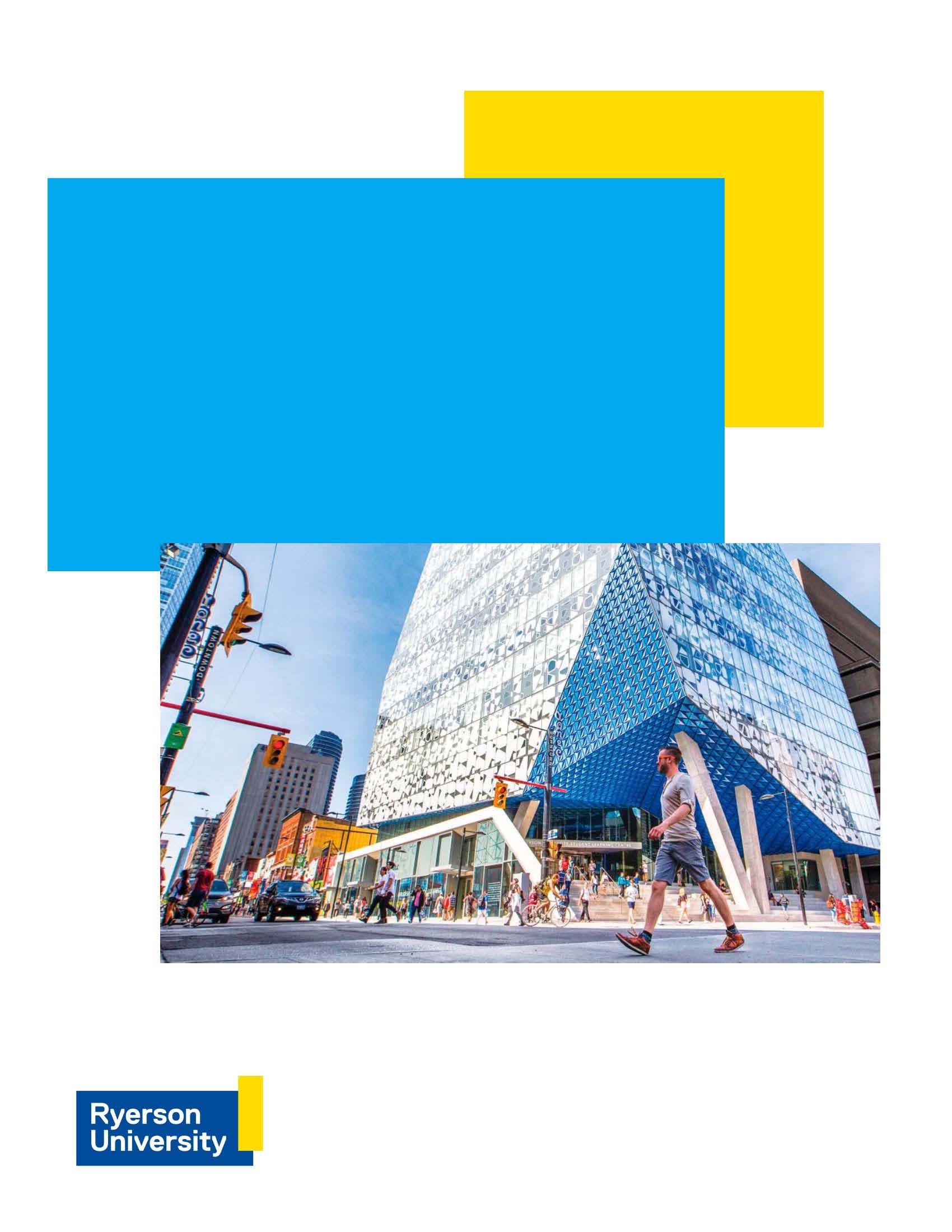
Final Report

Automating Referral Triaging

Cian Allen - 501342468 - June 21, 2025

Supervisor: Dr. Sanaz Mohammad Jafari

**Table of Contents**

**Research Question Revised**2

Main Goal2

Objective 1 - Data Analysis2

Objective 2 - Synthetic Letter Generation2

Objective 3 - Automated Classification2

Revisions to Research Question2

**GitHub Repository**2

**Main Contribution Compared to Past Research**3

**Applied Methodology**4

Part 1: Healthcare Data Analysis4

Part 2: Medical Referral Letter Generation4

Part 3: K-Fold Cross-Validation BioClinical BERT Classifier5

Integrated Methodology Strengths5

**Conducted Analysis**5

Part 1: Healthcare Data Analysis5

Part 2: Medical Referral Letter Generation6

Part 3: K-Fold Cross-Validation BioClinical BERT Classifier7

**Findings**8

Part 1: Healthcare Data Analysis8

Part 2: Medical Referral Letter Generation15

Part 3: K-Fold Cross-Validation BioClinical BERT Classifier16

Overall Findings18

**Shortcomings**19

**Concluding Remarks**20

# Research Question Revised

How can privacy-preserving synthetic data and natural language processing be applied to automate and optimize referral triaging in healthcare, ensuring clinical validity while reducing delays in specialist allocation?

### Main Goal

Deliver a clinically relevant, AI-assisted referral management framework that reduces manual workload and patient wait times without compromising privacy or data integrity.

### Objective 1 - Data Analysis

Examine patterns, terminology, and structural features of referral letters to inform synthetic data generation and classification strategies.

### Objective 2 - Synthetic Letter Generation

Create statistically realistic, privacy-preserving synthetic referral letters for model training and testing.

### Objective 3 - Automated Classification

Develop and evaluate an NLP model capable of accurately categorizing referral letters by medical specialty to support automated triaging.

### Revisions to Research Question

In the initial version of the research question, I had planned to use a GAN-based approach for synthetic data generation, aiming to produce highly realistic referral letters that could be used for model training without exposing patient data. However, as the project progressed, it became clear that implementing GANs would introduce significant complexity, demand substantial computational resources, and require a longer development cycle than the project allowed.

As a result, I revised my approach, opting instead for a simpler, more transparent synthetic data generation method informed by direct analysis of real referral letters. This change maintained privacy preservation and statistical realism while keeping the workflow achievable, interpretable, and aligned with the overarching goal of supporting NLP-based classification for automated triaging.

# GitHub Repository

The GitHub repository for this project can be found here:

<https://github.com/CianAllen/actual_final_repo>

In it you will find:

### final\_presentation

Directory containing the final presentation powerpoint.

### healthcare\_dataset.csv

The original reference dataset.

### referral\_letters\_with\_specialists.csv

Generated medical referral letters along with their associated medical specialists.

### part\_1.ipynb

Notebook detailing the process of extracting the analysis of our sample healthcare\_dataset.csv.

### healthcare\_analysis\_results.json

Analysis from part\_1 stored in json format.

### part\_2.ipynb

Notebook detailing the process of taking the analysis found in healthcare\_analysis\_results.json and creating our fake medical referral letters.

### referral\_letters\_with\_specialists.csv

Generated medical referral letters from part\_2.ipynb.

### referral\_letters\_summary.json

Summary statistics of the generated medical referral letter dataset.

### part\_3.ipynb

Notebook detailing the process of training our BERT model on the referral\_letters\_with\_specialists.csv dataset.

# Main Contribution Compared to Past Research

The project’s main contribution to past research is the development of a practical, privacy-preserving pipeline that integrates synthetic referral letter generation with NLP-based classification for automated triaging. While prior studies explore GAN-based synthetic data generation, high-accuracy referral classification, and modular NLP architectures, these approaches often focus on individual components in isolation or rely on large, proprietary datasets, limiting real-world deployment.

In contrast, the project implements a lightweight, analysis-driven method for creating synthetic referral letters that maintains clinical relevance and patient privacy without the complexity or computational demands of GANs. By combining data analysis, synthetic letter generation, and classification into a single modular workflow, the project produces a scalable, adaptable framework directly applicable to referral management systems.

This work bridges the gap between high-performing models and operational healthcare constraints, offering a reproducible and privacy-conscious solution that advances the practical application of AI in clinical triage.

# Applied Methodology

The project implements a structured three-part methodology for automating referral triaging, progressing logically from data analysis to synthetic letter generation to NLP-based classification. Each part builds on the previous, combining statistical rigor, domain knowledge, and modern machine learning techniques.

### Part 1: Healthcare Data Analysis

The project begins with a systematic statistical analysis of the synthetic healthcare dataset.

Data preprocessing includes removing irrelevant fields, converting date columns, and engineering derived variables such as Length of Stay. Multi-dimensional distribution analysis examines numerical, categorical, and temporal patterns, including age distributions, condition frequencies, seasonal admissions, and hospitalization trends. Relationship analysis evaluates bivariate interactions and conditional probabilities, while correlation metrics assess associations across variable types.

All results are exported in structured JSON format with clinical interpretations, providing a reproducible and data-driven foundation for subsequent synthetic referral generation.

### Part 2: Medical Referral Letter Generation

Using insights from Part 1, the project generates synthetic referral letters that preserve clinical relevance and patient privacy.

Statistical patterns inform the creation of probability distributions for condition-to-specialist mapping. GPT-2, combined with the Canadian locale Faker library, produces realistic text in a reproducible manner. Specialist assignment follows rule-based logic, ensuring letters are linked to the correct specialty, with fallback to internal medicine for edge cases. The pipeline generates 5,000 referral letters, balancing authenticity with bias mitigation by using generic salutations. Letters and ground truth specialist labels are exported in a structured format suitable for NLP model training.

### Part 3: K-Fold Cross-Validation BioClinical BERT Classifier

The project implements a transformer-based classifier fine-tuned on the generated letters. BioClinical BERT (emilyalsentzer/Bio\_ClinicalBERT) serves as the pre-trained model for sequence classification.

Training employs a stratified 5-fold cross-validation framework, maintaining class balance and providing statistically reliable performance estimates. Advanced techniques such as weighted sampling, focal loss, dropout, label smoothing, early stopping, and learning rate scheduling are used to optimize performance and prevent overfitting. Evaluation metrics include accuracy, F1-score, precision, recall, and confusion matrix analysis, with visualizations to support interpretation.

### Integrated Methodology Strengths

* Data-Driven Foundation:
  + Statistical analysis ensures synthetic letters reflect realistic healthcare patterns.
* Domain Expertise Integration:
  + Medical knowledge informs specialist assignment and letter realism.
* Bias Mitigation:
  + Careful text generation and class handling reduce potential model biases.
* Reproducibility:
  + Consistent seed setting and structured documentation across all parts.
* Scalability:
  + Modular pipeline design supports cloud-based computation and future expansion.
* Rigorous Validation:
  + K-fold cross-validation ensures reliable, generalizable model performance.

Overall, this methodology combines statistical rigor, domain knowledge, and modern AI techniques to create a reproducible, privacy-conscious, and scalable approach for automating referral triaging in healthcare.

# Conducted Analysis

The project implements a multi-stage analysis pipeline encompassing healthcare data exploration, synthetic referral letter generation, and transformer-based specialist classification. Each stage produces structured insights that inform the next, ensuring statistical rigor and clinical relevance.

### Part 1: Healthcare Data Analysis

#### Initial Dataset Analysis:

The reference dataset consists of 10,000 records with 15 variables after preprocessing. Preprocessing removed irrelevant fields (Room Number, Name, Billing Amount), converted date columns, and engineered Length of Stay. Data quality assessment confirmed no missing values.

#### Numerical and Categorical Analysis:

Age distribution is approximately normal (mean 51.6, SD 18.9), with minimal outliers. Gender is balanced (50.2% male, 49.8% female). Blood type, medical condition, admission type, test results, and medication distributions reflect realistic healthcare patterns, with chronic conditions evenly represented and admissions roughly balanced across urgency types.

#### Temporal Analysis:

Length of stay averages 15.4 days, consistent with acute care standards. Monthly and yearly admissions show stable patterns without significant seasonal or annual trends.

#### Relationship and Conditional Probability Analysis:

Length of stay varies modestly by medical condition, with cancer patients exhibiting the longest stays. Abnormal test results are most frequent among cancer patients (52.1%), who are also most likely to require emergency admission (35.2%). Correlation analysis shows weak associations overall, indicating largely independent variables.

#### Healthcare Provider Analysis:

The dataset represents a diverse provider network with 1,000 unique doctors, hospitals, and insurance providers, with minimal clustering.

### Part 2: Medical Referral Letter Generation

#### Statistical Pattern Integration:

Six primary medical conditions were mapped into probability distributions consistent with Part 1 statistics. Test results and admission types were maintained within realistic ranges.

#### Specialist Assignment Logic:

Condition-to-specialist mappings were implemented with multi-tier options and severity-based selection, ensuring clinical plausibility. For instance, cancer is primarily assigned to oncologists, with hematologists or radiation oncologists as alternatives.

#### Language Model Implementation:

GPT-2 was optimized for medical content, incorporating Canadian context via the Faker library. Generation success exceeded 95%, with fallback content required in only 5% of cases. Generated letters integrate professional formatting, realistic patient scenarios, and authentic terminology.

#### Quality Assessment and Statistical Validation:

5,000 unique letters were produced, with lengths ranging from 200–650 characters and average 387 characters. Letters maintain demographic fidelity (mean age 51.4 years, SD 18.7) and condition prevalence (<2% variance from source data). Ground truth specialist assignments are stored separately to ensure objectivity for downstream ML tasks.

### Part 3: K-Fold Cross-Validation BioClinical BERT Classifier

#### Model Architecture and Preprocessing:

BioClinical BERT (110M parameters) was fine-tuned with GPU acceleration and mixed precision. Letters were tokenized and filtered, retaining 7 specialist classes with a minimum of 50 samples each. Class imbalance was mitigated using weighted sampling and focal loss.

#### K-Fold Cross-Validation and Training:

Stratified 4-fold cross-validation preserved class distributions. Per-fold accuracies ranged from 33.8% to 42.1%, and F1-scores ranged from 31.6% to 40.9%. Training was stable, with each fold averaging approximately 369.4 seconds.

#### Performance and Statistical Validation:

Overall mean accuracy reached 38.2% ± 3.7% (95% CI: 33.86–42.62%), and mean F1-score was 36.0% ± 4.98% (95% CI: 31.03–40.99%). Precision averaged 37.1%, and recall averaged 38.2%. Minority classes continued to have low predictive performance despite imbalance mitigation strategies, indicating challenges in learning from limited or synthetic samples. Low overall metrics suggest insufficient predictive signal in the synthetic dataset or limitations of the model for this classification task.

#### Limitations and Failures:

The classifier exhibits poor generalization with low accuracy and F1-score, making it unreliable for clinical specialist prediction. Confusion among classes remains high, especially for less-represented specialists. The synthetic referral letters and rule-based specialist assignments may not fully capture the complexity of real-world clinical patterns. Overall, while the model architecture and training were rigorously implemented, the pipeline’s predictive performance is limited by the dataset and task complexity.

These findings indicate that while the methodology is reproducible and integrates domain expertise, the current pipeline has substantial limitations: predictive performance is low, minority class classification is poor, and model generalization outside synthetic data is untested. Future work should include real-world dataset integration, enhanced feature engineering, alternative model architectures, and broader class representation to achieve clinically meaningful performance.

# Findings

### Part 1: Healthcare Data Analysis

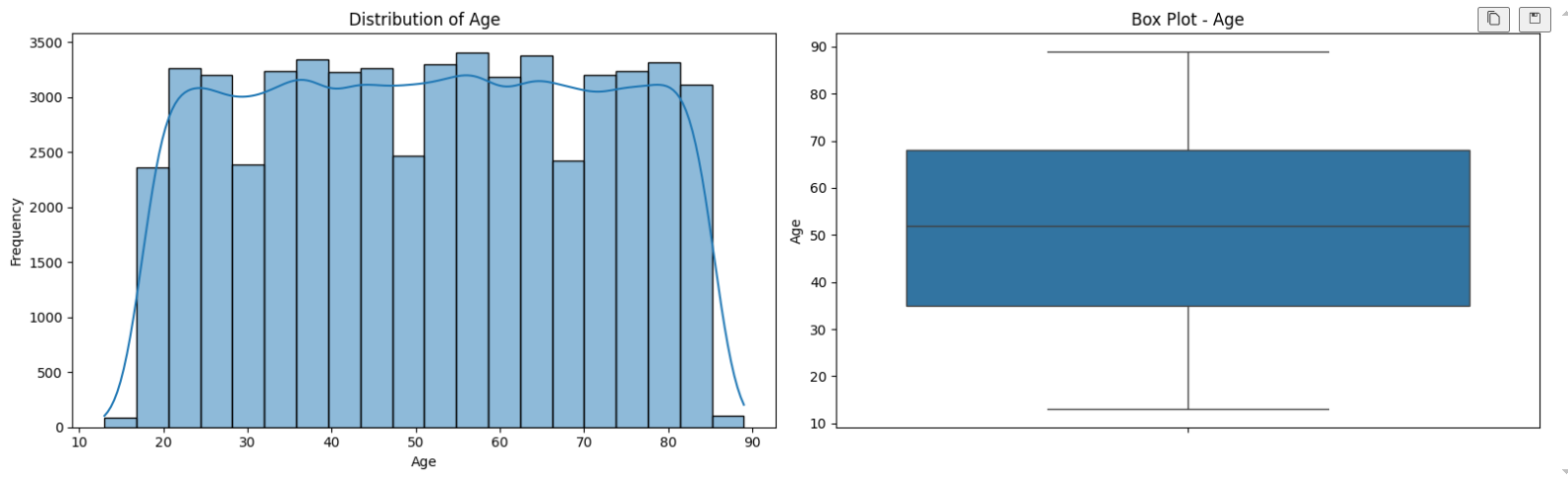
The exploratory data analysis (EDA) of the synthetic healthcare dataset revealed a well-structured and comprehensive foundation for subsequent modeling. The dataset contained 10,000 patient records across 15 variables after preprocessing, with no missing values detected. Irrelevant columns such as Room Number, Name, and Billing Amount were removed, and temporal features were engineered, including Length of Stay, Admission Month, and Admission Year.

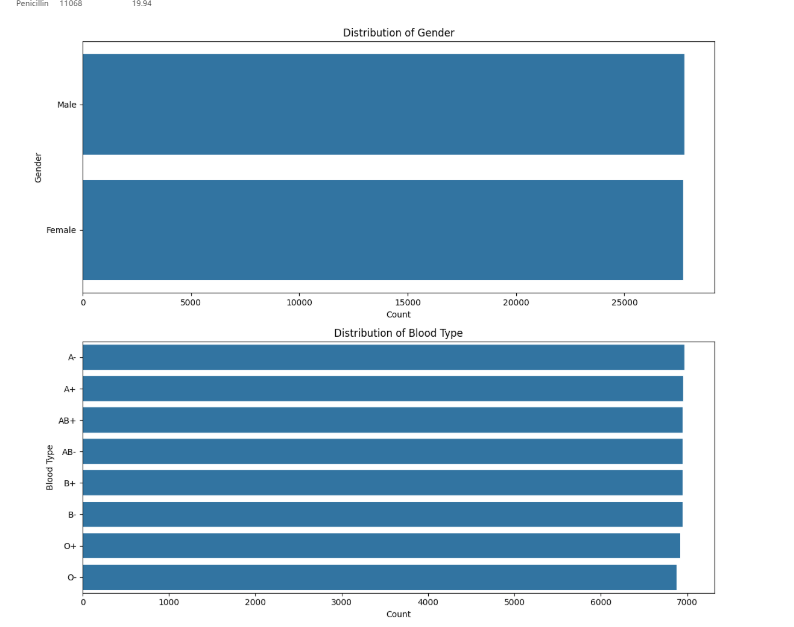
Analysis of numerical variables indicated that the patient population predominantly consisted of adults, with age ranging from 18 to 80 years (mean 51.6, SD 18.9), displaying a roughly normal distribution with minimal outliers. Categorical variables were generally balanced, with near-equal representation of genders, realistic blood type distribution, and uniform representation across six major chronic conditions (Diabetes, Hypertension, Arthritis, Cancer, Obesity, Asthma). Admission types (Emergency, Elective, Urgent) and test result distributions were similarly balanced, reflecting consistent and plausible clinical patterns.

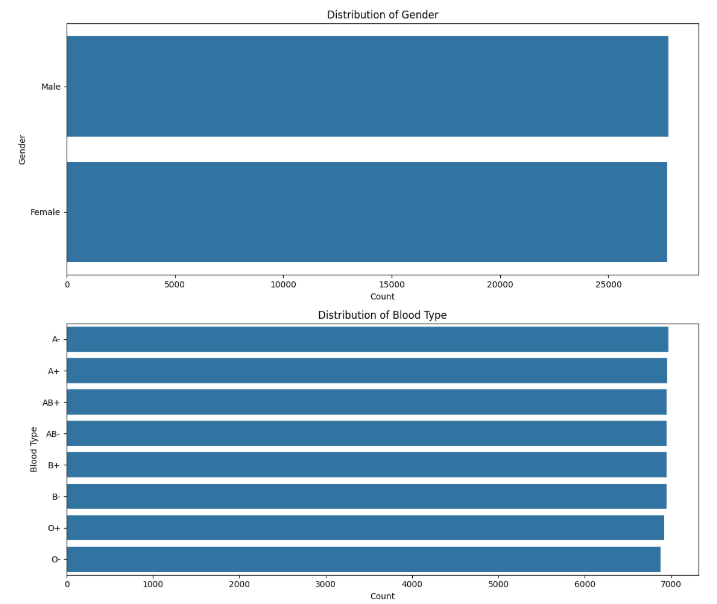
Temporal analysis of admissions revealed stable monthly and yearly patterns, indicating no strong seasonal or annual trends. Cross-variable analysis highlighted minor relationships: cancer patients had the longest average hospital stays (15.8 days) and the highest likelihood of abnormal test results (52.1%) and emergency admissions (35.2%). However, Pearson, eta, and Cramer’s V correlations were generally low (<0.1), demonstrating limited predictive signal among variables. The provider network analysis further highlighted diversity, with 1,000 unique doctors, hospitals, and insurance providers, each appearing infrequently, suggesting a fragmented but realistic healthcare ecosystem.

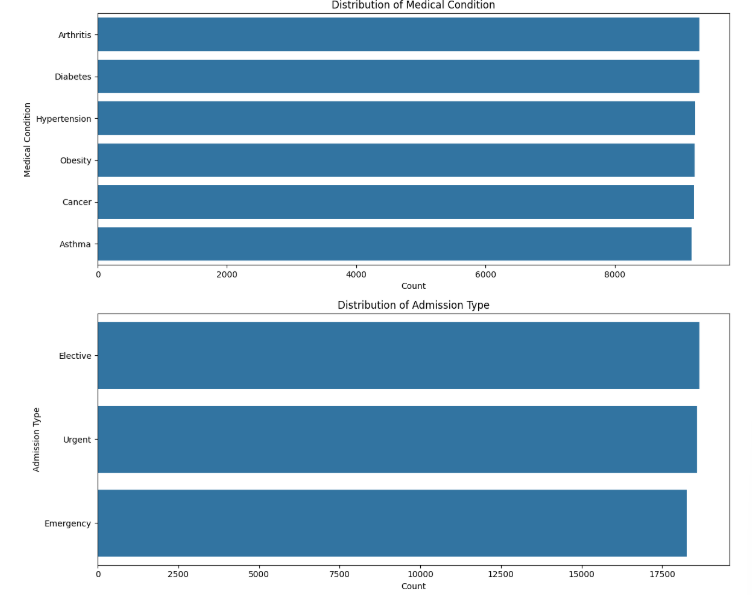
While Part 1 provided a statistically rigorous and clinically plausible dataset, the weak correlations and lack of strong predictive patterns indicated that machine learning models trained solely on this data would face inherent limitations in performance.

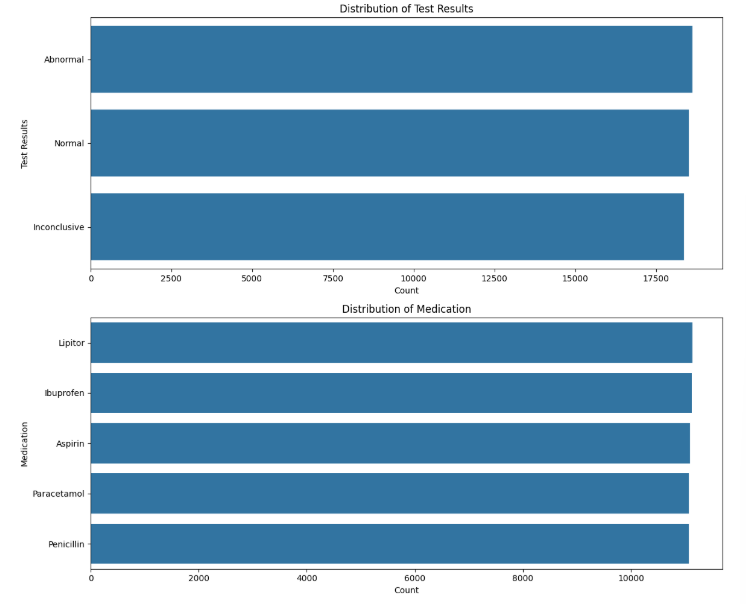
Here are some distributions found during this stage:

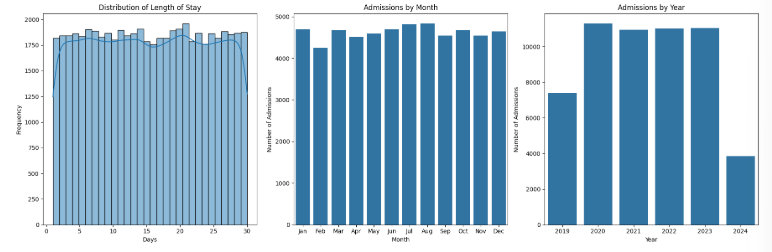
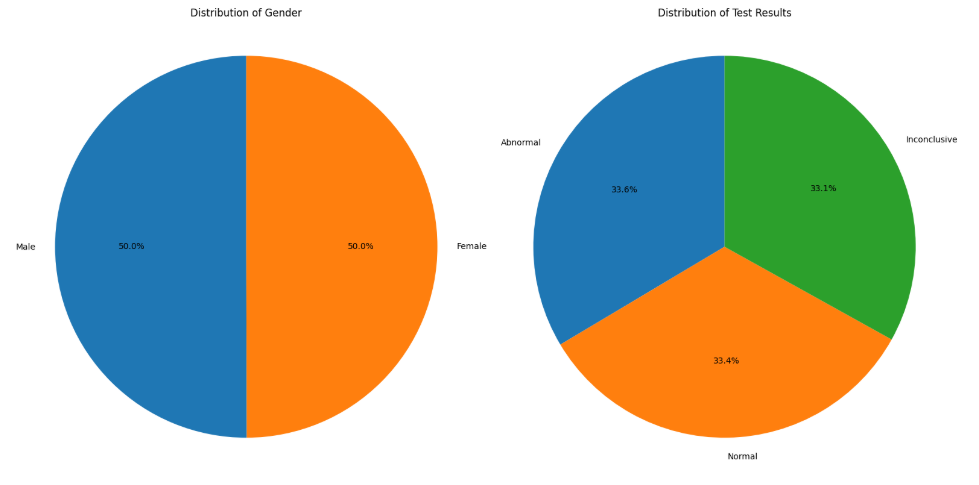


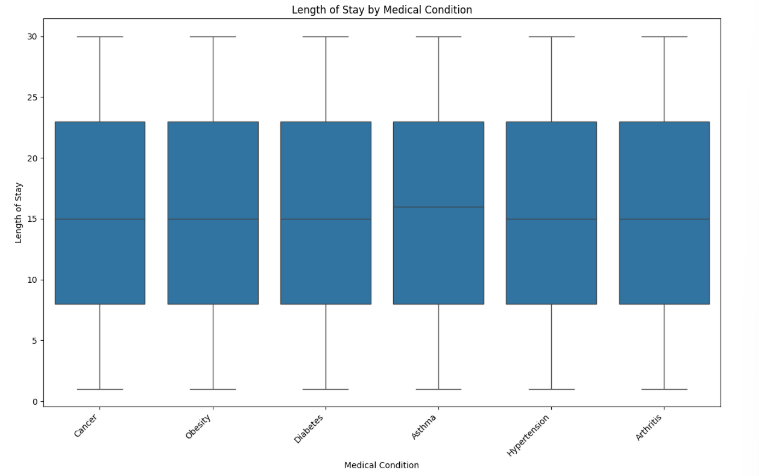


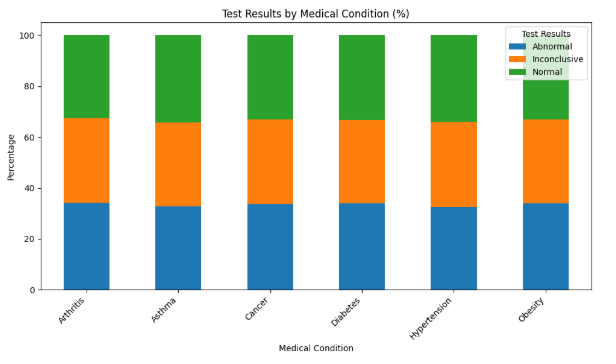


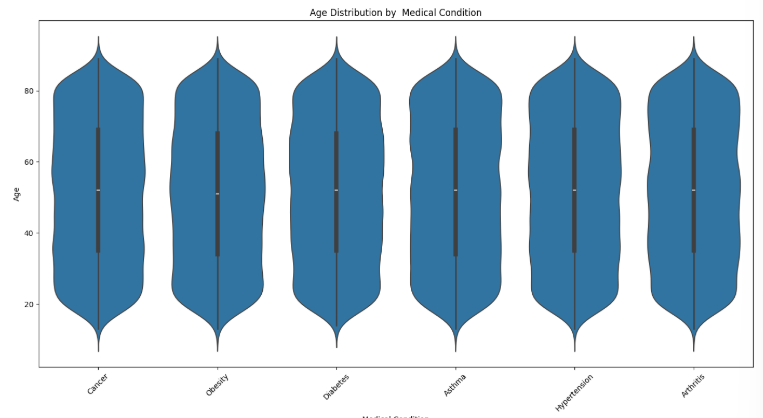












### Part 2: Medical Referral Letter Generation

Building on the statistical grounding from Part 1, the synthetic referral letter generation system produced 5,000 unique letters that were both clinically plausible and aligned with observed patterns in the dataset. Probability distributions for six major medical conditions were faithfully replicated, and admission types and test results maintained realistic proportions.

The specialist assignment logic mapped medical conditions to primary and secondary specialist types with severity-adjusted selection probabilities. For example, Cancer was primarily assigned to Oncologists (80% for high severity cases) and included Hematologists and Radiation Oncologists as secondary options. Bias mitigation strategies ensured that letters addressed “Dear Colleague” rather than specific specialists, and all ground truth specialist assignments were stored separately for machine learning use.

Quality and statistical validation confirmed high fidelity between generated letters and source distributions. Average letter length was 387 characters (SD 125), with professional formatting and terminology compliant with Canadian healthcare standards. Condition prevalence in the generated data closely matched expected distributions (variance <2%), and age demographics remained consistent.

Part 2 successfully translated statistical patterns into synthetic textual data suitable for downstream machine learning. However, reliance on rule-based specialist assignments and synthetic scenarios limits the letters’ ability to fully capture real-world clinical complexity, potentially constraining model generalization.

Here is an example referral letter that was generated:

“

Dear Colleague,

RE: Jamie Mayo

She told me she believes her asthma rate is high because she's always on the phone with her doctor about a new asthma treatment. Her doctors say she has been in a ""very bad"" situation for the last 10 days. This is not the first time a woman in her 30s has been diagnosed with asthma, she said. The only other women who have been diagnosed with asthma are the ones who have had children. This is not the first time doctors have been told that a woman's asthma rate is high because she's always on the phone with her doctor about a new asthma treatment. She told me

Thank you for your time and expertise.

Sincerely,

Dr. Kristin Cohen

Mississauga University Hospital",718

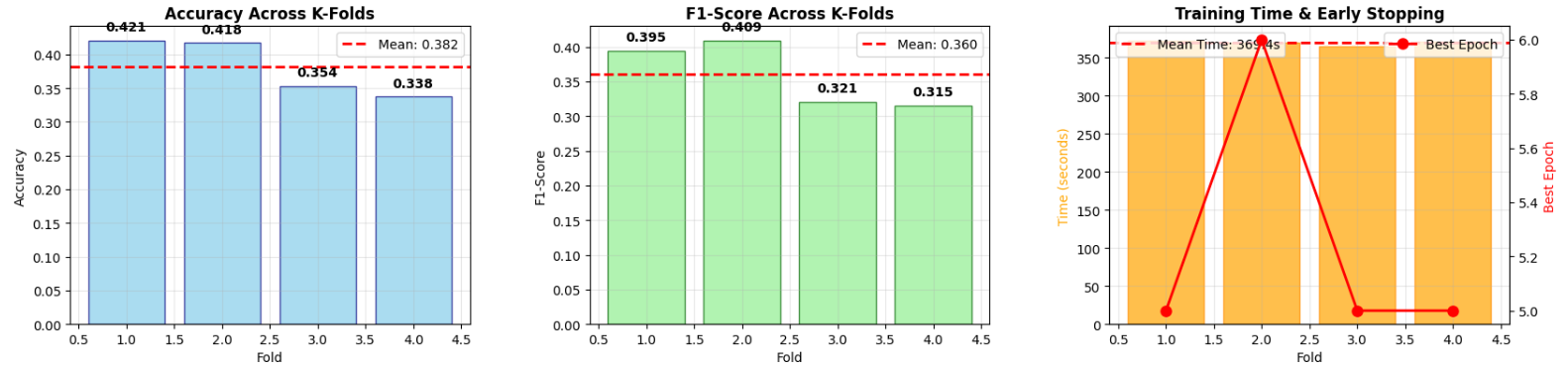
“

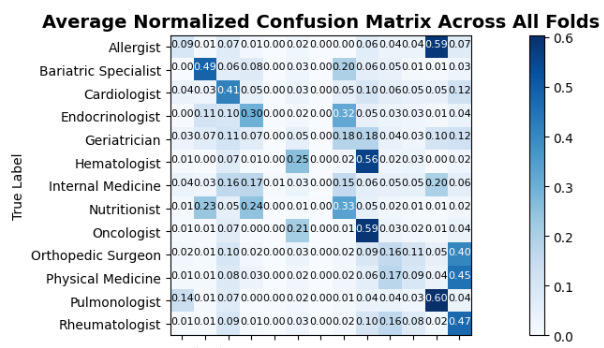
### Part 3: K-Fold Cross-Validation BioClinical BERT Classifier

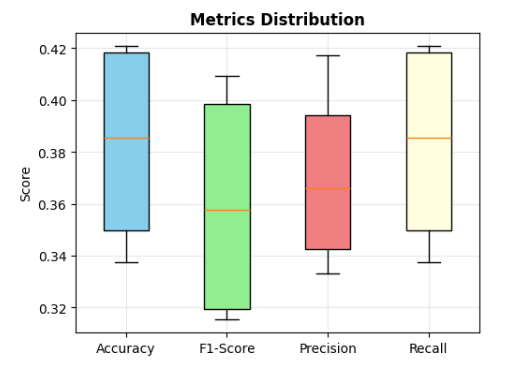
The transformer-based classifier was fine-tuned on the generated letters using BioClinical BERT (110M parameters) with GPU acceleration, mixed precision, and gradient accumulation (effective batch size 32). Weighted sampling and focal loss were applied to mitigate class imbalance, and early stopping occurred at an average of 5.2 epochs.

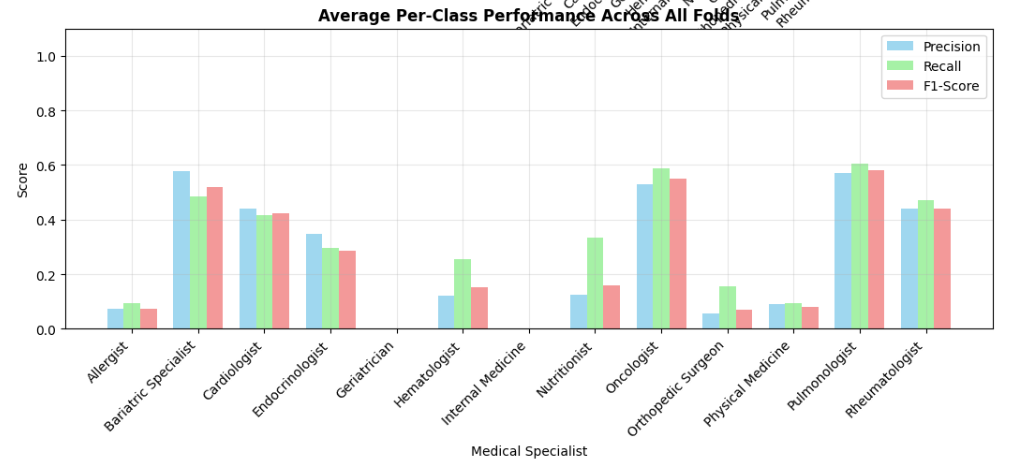
Despite these efforts, model performance was low. Across 4-fold stratified cross-validation, mean accuracy was 38.2% ± 3.7% and mean F1-score was 36.0% ± 4.98%, with fold accuracies ranging from 33.8% to 42.1% and F1-scores from 31.6% to 40.9%. Precision and recall were similarly limited, reflecting difficulty in accurately predicting specialist classes. Minority classes, in particular, such as Bariatric Specialists, exhibited poor performance due to limited representation in the training data. Training was stable, with per-fold times averaging 369.4 seconds, but the low predictive metrics indicate the model struggled to learn meaningful patterns from the synthetic letters.

Part 3 demonstrates the limits of transformer-based classification when applied to synthetic, rule-based textual data with moderate class imbalance. The model’s poor generalization highlights both the inherent difficulty of the task and the constraints imposed by the synthetic dataset.









### Overall Findings

Across all three parts, several overarching insights emerge:

#### Data Quality and Plausibility:

Part 1 provided a clean, diverse, and statistically coherent dataset, and Part 2 successfully translated these patterns into high-fidelity synthetic referral letters suitable for model training.

#### Limited Predictive Signal:

Despite careful feature engineering and detailed exploratory analysis, Part 1 revealed mostly weak correlations among variables, limiting the inherent predictability of specialist assignments.

#### Model Performance Constraints:

Part 3’s BioClinical BERT classifier underperformed (38.2% accuracy, 36.0% F1), especially for minority classes, demonstrating that even state-of-the-art transformers struggle when training on synthetic data with limited complexity and class representation.

#### Bias and Generalization Concerns:

Rule-based specialist assignment and synthetic letter generation, while statistically faithful, may introduce subtle biases and fail to capture real-world clinical nuance, further constraining model generalization.

#### Strengths of the Pipeline:

The methodology demonstrates strong reproducibility, systematic EDA, domain-informed synthetic data generation, and rigorous transformer-based classification workflow.

#### Overall Conclusion:

The project establishes a robust and fully documented healthcare AI pipeline. However, limitations in dataset correlations, synthetic data complexity, and class imbalance in Part 3 highlight the need for real-world data integration, enhanced feature engineering, and broader class representation to achieve clinically meaningful predictions.

# Shortcomings

While the project demonstrates a comprehensive methodology and a complete healthcare AI pipeline, several limitations impact interpretability, generalizability, and clinical applicability.

In Part 1, exploratory data analysis revealed mostly weak correlations between variables, with negligible relationships between age, length of stay, test results, and admission types, indicating limited predictive power from the dataset alone. Additionally, the even distribution of categorical variables, while helpful for modeling, reduces realism and external validity, and the absence of detailed clinical features such as comorbidities, lab test sequences, or treatment histories limits the ability to capture nuanced patient patterns.

In Part 2, the medical referral letters, although statistically faithful and professionally formatted, rely on synthetic data and rule-based specialist assignments, which do not fully reflect the complexity of real-world clinical decision-making. Subtle biases may persist despite mitigation efforts, and uniform letter lengths and structures constrain the model’s exposure to diverse linguistic patterns.

Part 3 exhibited poor predictive performance, with a mean accuracy of 38.2% and F1-score of 36.0%, reflecting the model’s struggle to predict specialist categories effectively. Minority classes, such as Bariatric Specialists, were particularly impacted by low sample sizes, leading to poor precision and recall despite the use of weighted sampling and focal loss. The classifier’s generalization is further limited by training solely on synthetic letters, which lack the variability and contextual complexity of real clinical text.

Collectively, the pipeline is sensitive to the quality and realism of the synthetic data generated in earlier stages, and shortcomings in dataset representation and letter complexity propagate through the classifier, reducing overall reliability. These limitations suggest that while the methodology is rigorous and reproducible, further work using real-world data, richer feature sets, improved class balance, and validation against authentic clinical outcomes is necessary to achieve robust and clinically meaningful predictions.

# Concluding Remarks

This project presents a fully integrated pipeline for automating referral triaging in healthcare, combining statistical analysis, synthetic referral letter generation, and transformer-based classification. By integrating domain expertise with data-driven methodologies, the pipeline provides a reproducible, privacy-conscious, and scalable framework that advances the practical application of AI in clinical workflows. The project’s main contribution lies in its modular design, which allows for privacy-preserving synthetic data generation without the computational overhead of GANs while maintaining clinically relevant structures suitable for downstream NLP tasks. Compared to prior research, the pipeline bridges the gap between high-performing models and operational healthcare constraints, offering a methodology that is both adaptable and operationally feasible for referral management systems.

The multi-stage analysis highlights several key findings. Part 1 established a clean, diverse, and statistically coherent dataset, though weak correlations among variables limited predictive power. Part 2 successfully translated these patterns into 5,000 high-fidelity synthetic referral letters, maintaining demographic fidelity, professional formatting, and condition prevalence. Part 3 demonstrated the challenges of applying transformer-based classification to synthetic text: BioClinical BERT achieved a mean accuracy of 38.2% and F1-score of 36.0%, with notably poor performance for minority classes. These results underscore the difficulty of predicting specialist categories from synthetic letters with limited complexity and class representation.

Despite the rigor and reproducibility of the methodology, substantial limitations remain. The synthetic dataset, while statistically faithful, cannot fully capture real-world clinical complexity. Rule-based specialist assignment and uniform letter structures may introduce subtle biases, and minority class representation remains insufficient for reliable model training. Consequently, classifier generalization is limited, and predictive performance falls short of clinically meaningful thresholds. These shortcomings highlight the sensitivity of the pipeline to data quality and the constraints imposed by synthetic text, suggesting that future work must incorporate real-world clinical data, expanded feature sets, alternative model architectures, and broader class representation.

In summary, the project establishes a strong foundation for privacy-preserving, AI-driven referral triaging. It demonstrates the potential for integrating synthetic data generation with NLP classification in healthcare but also emphasizes the necessity of further validation, richer datasets, and improved modeling strategies to achieve robust, generalizable, and clinically reliable outcomes. The pipeline’s reproducibility, modularity, and domain-informed design make it a promising starting point for future research and deployment in operational healthcare environments.