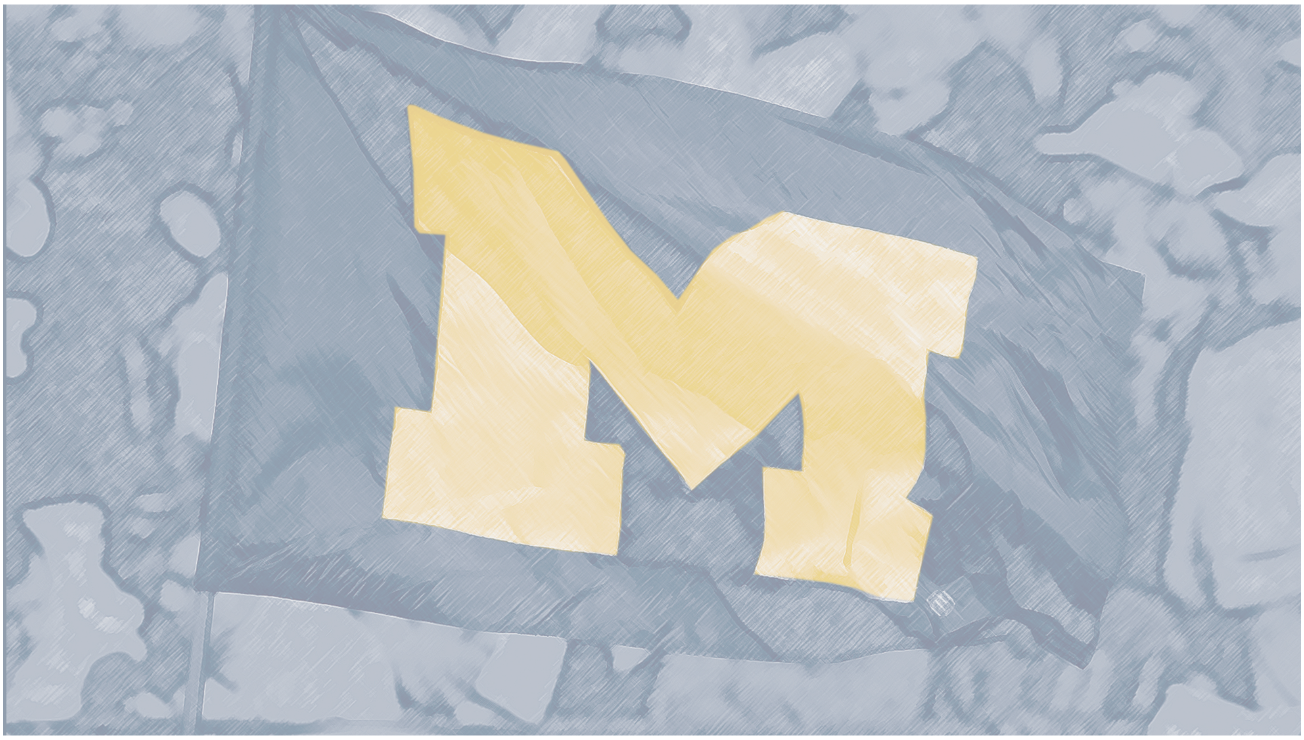


Prediction for potential medical treatment pathways with outpatient data

Capstone project

Team Waston

Xuan TAO, Jingyi LAI, Jeffery JIN

Introduction

Patients visiting hospital outpatient departments often face uncertainties about their diagnosis, necessary tests, and potential costs, which can be sources of significant distress. In the realm of healthcare, the integration of machine learning with clinical data has long been in practice, offering substantial benefits such as enhanced diagnostic accuracy and patient care optimization. While numerous studies have explored diagnostic and support pathways for specific ailments like depression and diabetes, the application of these technologies in creating interactive health self-check platforms remains less explored. This capstone project delves into this innovative area by posing below research questions:

i. Could analyzing medical encounter data reveal distinct patient groups with similar healthcare needs?
ii. Could classification prediction on user-reported symptoms and health status enable the identification of potential treatment pathways in a statistical representation?

Methods & Results

Dataset

We use the outpatient dataset of National Hospital Ambulatory Medical Care Survey (NHAMCS). An outpatient is a patient who receives medical treatment before or without necessitating hospital admission, often visiting a clinic or a separate department of a hospital for diagnosis or treatment. We focus on potential pathways and treatment insights, prioritizing managing patient expectations (see Figure 1).

Classification approach

Using ICD-9-CM's 'Classification of Diseases and Injuries', the model incorporates 17 labels to forecast healthcare needs based on demographics, prior visits, vital signs, and medical conditions. Figure 1 showcases the methodology of classifying patient data for predictive modeling. This classification approach underpins the development of a health self-check interactive platform, aiming to guide patients before professional consultations, with a focus on forecasting potential diagnostic or screening services.

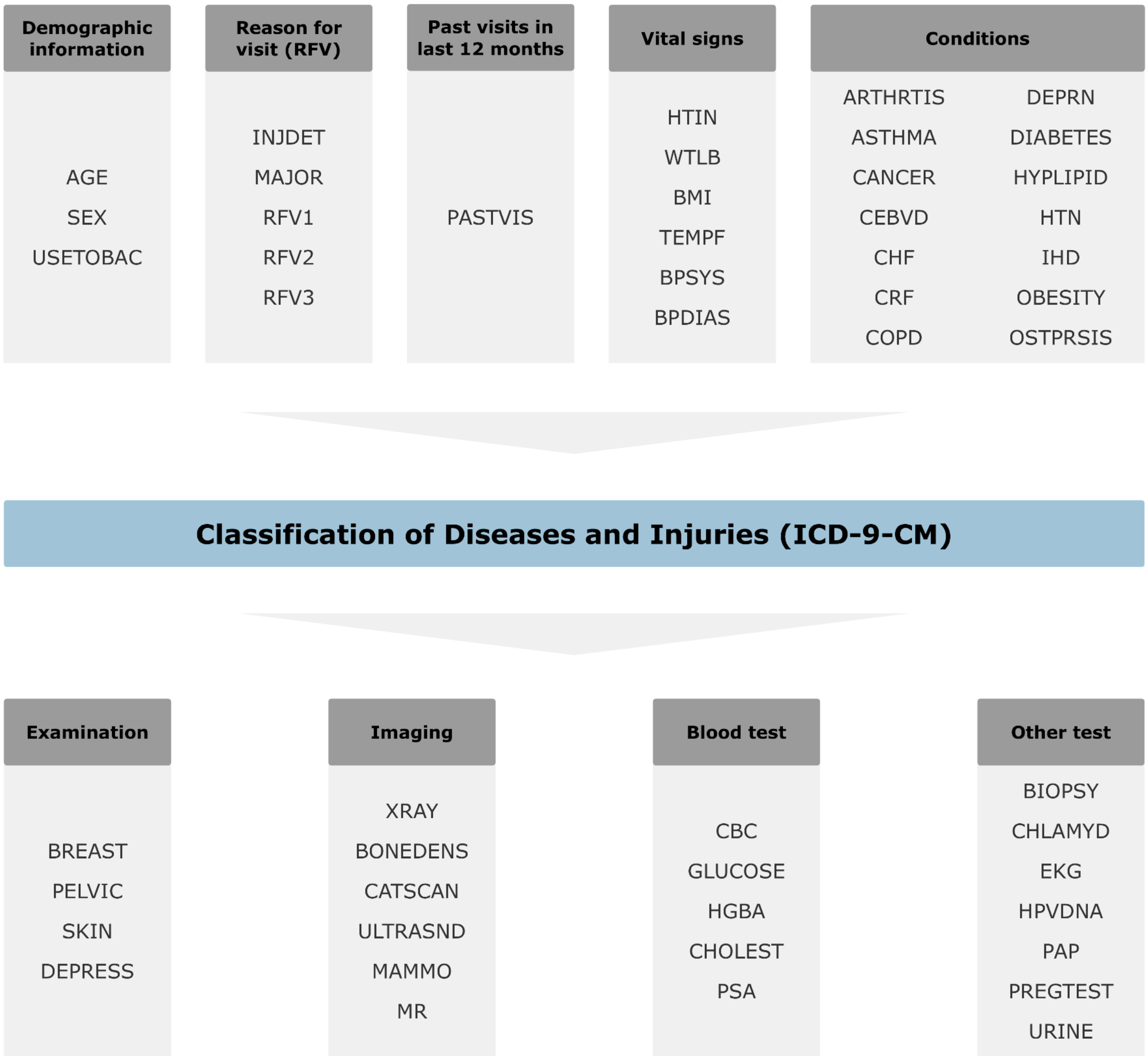


Figure 1: Structure of classification approach

Preprocessing

1. Merging

The ASCII outpatient data files, containing the original survey codes as values, were integrated and merged with text columns (e.g., REASON FOR VISIT, detailed in Appendix A Github Notebook 1.1-xt-merge_clean_raw_datasets.ipynb) from parsed outpatient data sets, using the SPSS code provided on the CDC website.

Discrepancies and inconsistencies are present in survey items across the years. For instance, from 2007 and on, the items "Is female patient pregnant, and, if so, specify gestation week." are removed. In 2009 and 2011, the cancer stage element was removed. Furthermore, from 2009, items for two scope procedures, one site of biopsy, two other diagnostic screening procedures, and four other surgical or non-surgical procedures are substituted with `PROC1` to `PROC9`. Thus, the outpatient data files for all six years have been merged by consolidating relevant columns across datasets.

2. Cleaning

Commencing in 2007, the National Center for Health Statistics (NCHS) implemented a standardized set of negative codes for representing missing data: -9 for 'Blank', -8 for 'Unknown', and -7 for 'Not applicable/calculated'. Prior to this, missing data was encoded using various positive numbers. While 'Blank' and 'Unknown' possess distinct meanings within the survey records, they represent an equivalent missing value status for the machine learning models in our analyses. 'Not applicable/calculated' may reflect diverse scenarios depending on the specific survey item. In order to streamline these situations, we have unified the codes for all missing data to -9. It is essential to acknowledge that treating all 'Not applicable/calculated' values as missing values may potentially exert an adverse effect on the models. Future research should investigate more nuanced approaches for handling missing data for different items.

3. Splitting

To ensure the generalizability and adaptability of data shift, we have decided to split the data in a chronological fashion (see Table 1).

Train	2006-2008	Approx. 51% of the total data
Validation	2009	Approx. 16% of the total data
Test	2010	Approx. 17% of the total data
Evaluation	2011	Approx. 16% of the total data

Table 1: Dataset split

Features engineering

In the Feature Engineering section, we turned raw data into a set of features, with which to prepare our data to make sure our models could make accurate predictions without being too complex. We then tested four different models to see which one predicted the best. We prepared each type of data using specific methods.

A. Management of missing data:

1. Categorical variables: missing values were treated as a separate category.
2. Quantitative variables: strategies were applied in accordance with the algorithms of different machine learning models, such as dropping columns, explicit

imputation, implicit model embedded binning strategy or more sophisticated decision rules.

B. We grouped continuous variables such as 'AGE' into clinically meaningful categories to improve model stability and reflect medical expertise (Table 2). Utilizing classification and hierarchical structure of REASON FOR VISIT (RFV) to add in domain expertise that beyond the original RFV codes (Table 3). The categorical variables received a meticulous makeover through One-Hot encoding, which expanded their representation for precise model ingestion.

AGE	Body Mass Index (BMI)	Temperature in Fahrenheit (TEMPF)	Blood Pressure - Systolic (BPSYS)	Blood Pressure – Diastolic (BPDIAS)
<20: Group 1 - Child or Teenager 20-40: Group 2 - Adult 40-60: Group 3 - Middle Aged >=60: Group 4- Senior	<18.5: Underweight 18.5-25: Normal weight 25-30: Overweight >=30: Obesity	<95: Hypothermia 95-99: Normal temperature 99-100: Low grade fever 100-103: Fever >=103: Hyperpyrexia	<90: Hypotension 90-120: Normal blood pressure 120-140: Prehypertension >=140: Hypertension	<60: Low diastolic blood pressure 60-90: Normal diastolic blood pressure 90-110: High diastolic blood pressure >=110: Hypertension

Table 2: Binning of Age, Body Mass Index, Temperature, Blood Pressure-Systolic, and Blood

RFV CODE RANGE	RFV_MODULE1	RFV_MODULE2
1001-1099	SYMPTOM MODULE	General Symptoms
1100-1199	SYMPTOM MODULE	Symptoms Referable to Psychological and Mental Disorders
…	…	…
2001-2099	DISEASE MODULE	Infective and Parasitic Diseases
2100-2199	DISEASE MODULE	Neoplasms
…	…	…
3100-3199	DIAGNOSTIC, SCREENING AND PREVENTIVE MODULE	General Examinations
3200-3299	DIAGNOSTIC, SCREENING AND PREVENTIVE MODULE	Special Examinations
…	…	…
4100-4199	TREATMENT MODULE	Medications
4200-4299	TREATMENT MODULE	Preoperative and Postoperative Care
…	…	…

Table 3: REASON FOR VISIT CLASSIFICATION (partially as an example)

C. As for the textual data, we conducted the following incorporation:

a) The initial model comparison reveals underfitting, evidenced by the similar accuracy of logistic regression, a simpler model, to that of more intricate ones like Random Forest. This suggests the model's complexity may not match the straightforward structure of our 29 categorical input features.

b) The categorical labels can be regarded as the abstracted high-level representation of natural phenomena (Figure 2). The level of detail of information decreases during the categorization process. By reversing the categorical labels to text descriptions, it is possible to recover the lost granularity. For example, the following sample illustrates how values are reversed to the text descriptions.

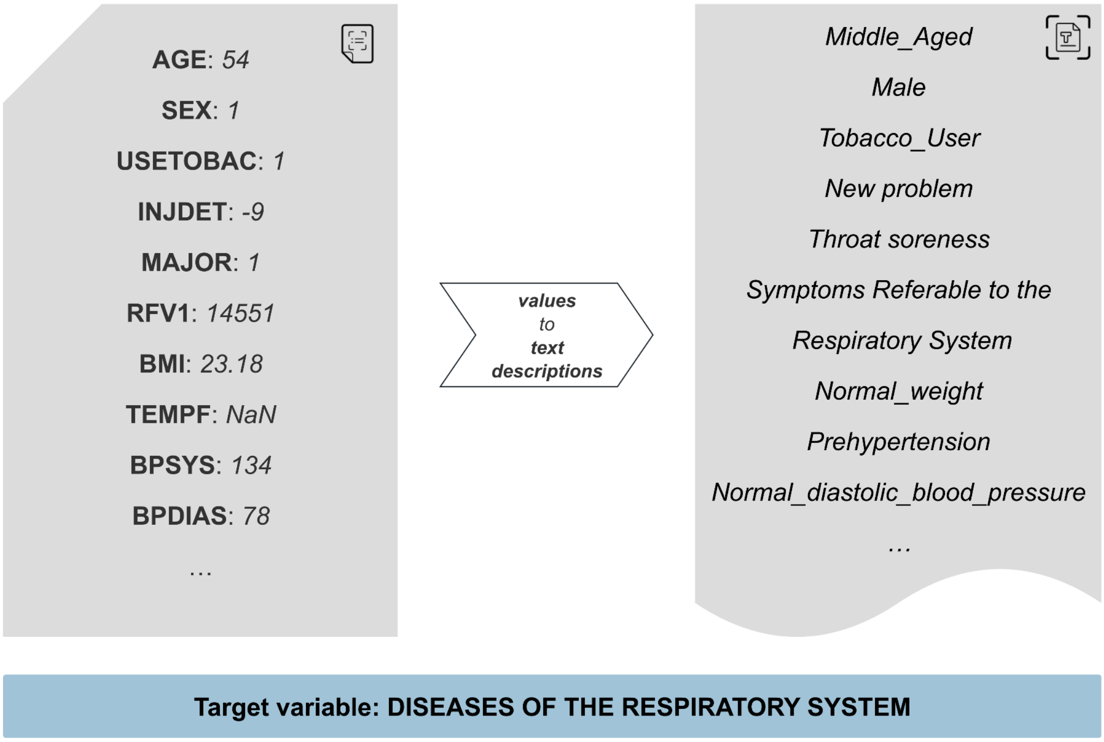


Figure 2: Illustration of values converted to text description for topic modeling

c) Textual data underwent lemmatization, with stop words and punctuation being removed, then transformed into numerical representations using TF-IDF scoring. Latent Dirichlet Allocation (LDA) was applied to generate topic features from these scores. This innovative approach of reversing categorical labels to text descriptions not only enhances the granularity of input features but also maintains a relatively low feature dimensionality, thereby mitigating the risk of overfitting and data sparsity. These topic features encapsulate the topic distributions of each combined textual description, with each topic being characterized by a distribution over words (Figure 3, see ./reports/figures/full_data_lda_vis.html in the GitHub repository for details). Figure 4 illustrates the likelihood of different topics across various classification of Diseases and Injuries identified in the full dataset. The heatmap visualization indicates the probability of each topic within a class, providing an at-a-glance view of prevalent medical issues associated with particular topics.

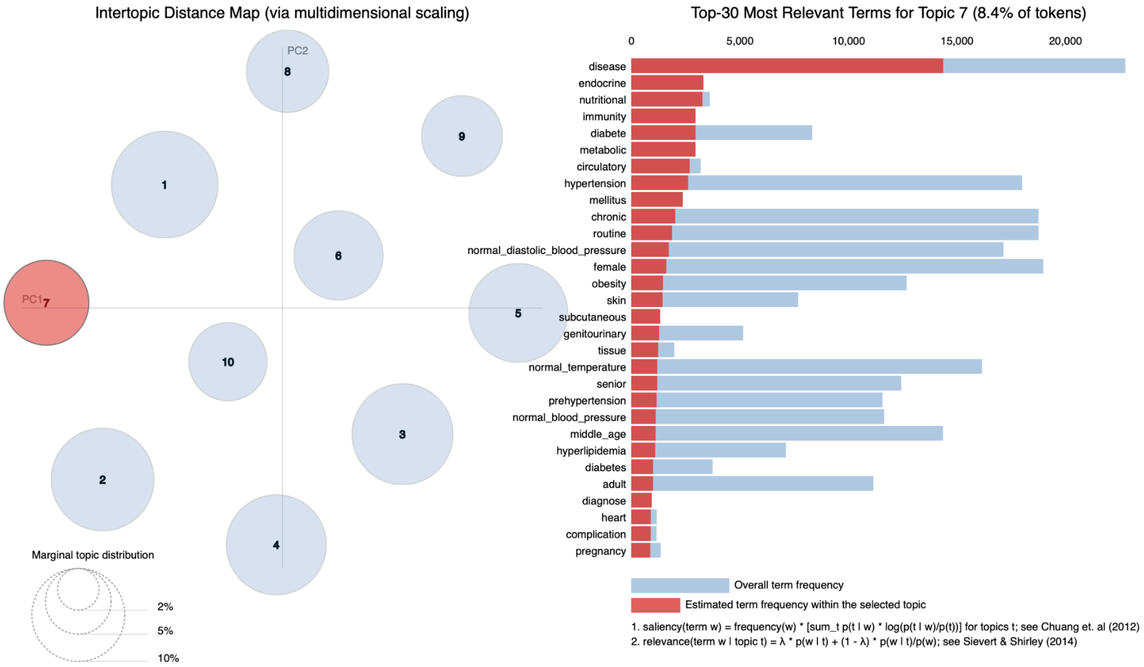


Figure 3: Intertopic Distance Map and Top-30 Most Relevant Terms for each topic (Topic 7 here as an example)

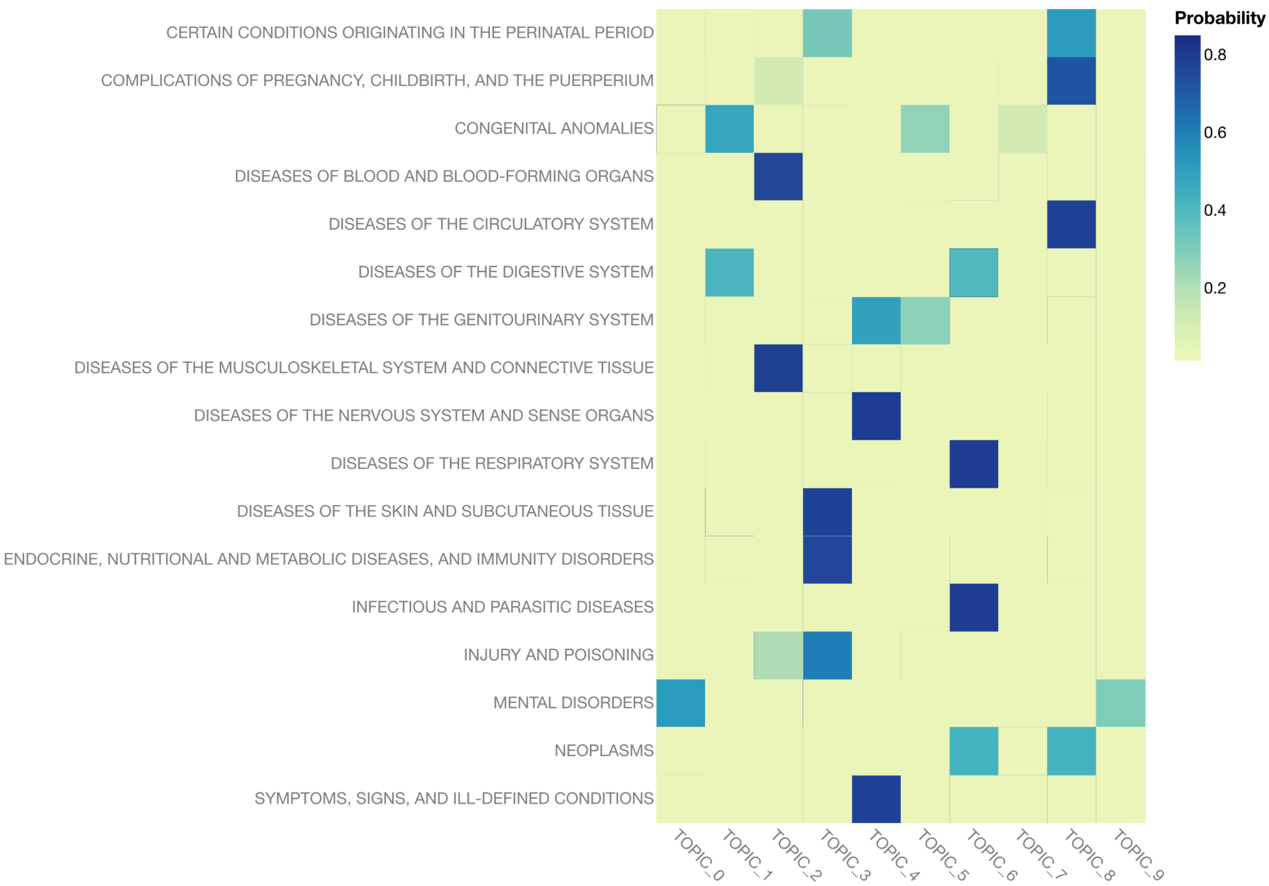


Figure 4: Distribution of the labels of each topic (with the full dataset)

Machine learning

We train the machine learning models with features listed on Table 4.

Categorical Features	SEX, USETOBAC, INJDET, MAJOR, ARTHRTIS, ASTHMA, CANCER, CEBVD, CHF, CRF, COPD, DEPRN, DIABETES, HYPLIPID, HTN, IHD, OBESITY, OSTPRSIS
Original Quantitative Features	AGE, PASTVIS, HTIN, WTLB, BMI, TEMPF, BPSYS, BPDIAS
Quantitative Features with Binned Groups	PASTVIS, HTIN, WTLB, AGE_GROUP, BMI_GROUP, TEMPF_GROUP, BPSYS_GROUP, BPDIAS_GROUP
RFV Codes	RFV1, RFV2, RFV3
RFV Modules	RFV1_MOD1, RFV2_MOD1, RFV3_MOD1, RFV1_MOD2, RFV2_MOD2, RFV3_MOD2
Topic Features	Converted and combined from AGE, AGE_GROUP, SEX, USETOBAC, INJDET, MAJOR, RFV1, RFV2, RFV3, BMI, TEMPF, BPSYS, BPDIAS and `CONDITIONS`

Table 4: Classification of Features Utilized in Machine Learning Model Analysis

A. Selection of viable machine learning models

1. **Logistic Regression Classifier:** Chosen for its speed and effectively handles the predominantly categorical data, this classifier yields interpretable probabilities, signifying class likelihoods essential for decision-making in our later analysis.
2. **Random Forest Classifier**: This classifier excels in handling non-linear feature-target relationships and is robust to overfitting, offering an edge over Logistic Regression, especially when dealing with missing values without relying on imputation.
3. **Histogram-Based Gradient Boosting Classifier (HGBC)**: HGBC employs a distinctive approach in handling missing data, treating them as a separate category or bin. The embedded algorithm utilizes histograms to bin the quantitative features which is different from our manual binning strategies. This binning process effectively condenses the number of trees, leading to a significantly smaller final model compared to the Random Forest Classifier (25 megabytes vs 10 gigabytes).

4. **Multi-Layer Perceptron Classifier (MLP)**: MLP's use of nonlinear activation functions pointed to its potential for modeling complex, nonlinear relationships within our data. Although it indicated a need for more sophisticated feature processing, such as BERT's domain-specific word embeddings to capture deeper semantic patterns, practical constraints prevented their implementation.

B. Evaluation metric (Weighted F1-score)

The weighted F1-score is essential for evaluating classifiers in **imbalanced** medical datasets, ensuring that both precision and recall are balanced across underrepresented categories, for a fair assessment of model performance (Figure 5).

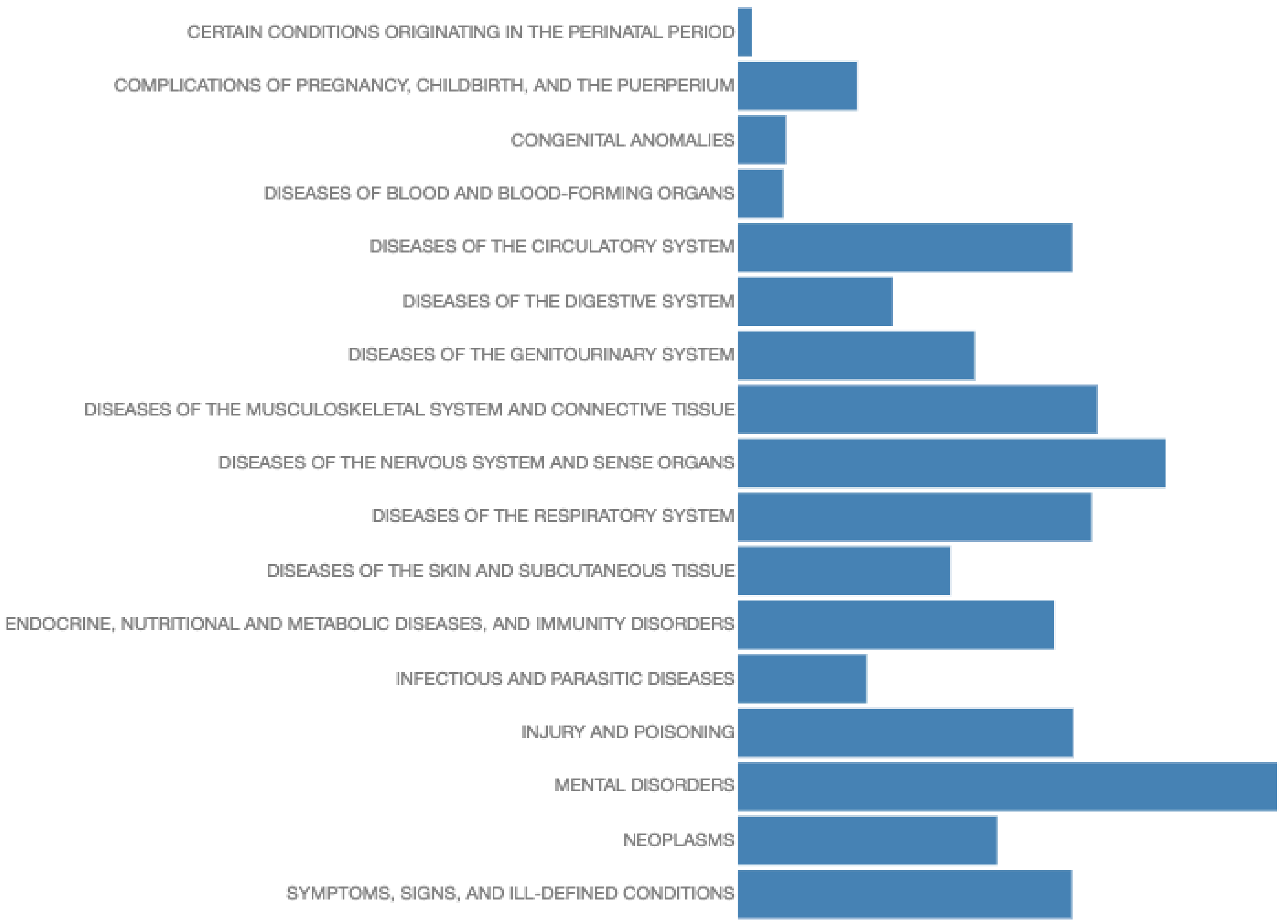


Figure 5: Imbalance distribution of target variables

C. Preliminary comparison of five models (Table 5)

1. We applied a 5-fold cross-validation procedure to evaluate the performance of various models constructed with different combinations of features using the training dataset. The **baseline model**, a logistic regression classifier model with only the 'CONDITIONS' features, achieved an average weighted F1-score of 0.2626. This score is notably higher than what would be expected by random chance (1/17 = 0.0588). This result suggests that the underlying logic of the

correlation and prediction holds and is not merely a product of chance.

2. For the **Logistic Regression Classifier model**, the combination of Quantitative Features with Binned Groups, RFV Codes, RFV Modules, and Topic Features, along with the use of a KNN Imputer with n_neighbors = 5 for handling missing values, resulted in the highest average weighted F1-score of 0.6141.
3. Varying combinations of features for the **Random Forest Classifier** yield a counter-intuitive outcome. Incorporating Topic Features results in a decline in performance metrics. This phenomenon suggests the potential existence of the "curse of dimensionality," requiring a larger dataset for effective model fitting.
4. The **Histogram-Based Gradient Boosting Classifier**, leveraging a blend of categorical, quantitative, RFV, and topic features along with KNN imputation, outperforms other models with a leading score of 0.6430, marking a 2.9% increase from its closest competitor.
5. The performance of the **Multi-Layer Perceptron Classifier** remains markedly inferior to that of alternative models, despite the incorporation of Topic Features. This suggests that the simplicity of the current input data does not align with the requirements of this model.

Model	Features	Management of missing data	Mean CV Score (Weighted F1)
Baseline: Logistic Regression Classifier	'CONDITION' only	-	0.2626
	Categorical Features + Quantitative Features with Binned Groups + RFV Modules	Dropping columns with missing values (PASTVIS, HTIN, WTLB)	0.6023
	Categorical Features + Original Quantitative Features + RFV Codes	Dropping columns with missing values (e.g., PASTVIS, HTIN, WTLB, BMI)	0.6079
	Categorical Features + Original Quantitative Features + RFV Codes	KNN Imputer (n_neighbors = 5)	0.6092
Logistic	Categorical Features + Quantitative Features	KNN Imputer (n_neighbors = 5)	0.6134

Regression Classifier	with Binned Groups + RFV Codes		
	Categorical Features + Quantitative Features with Binned Groups + RFV Codes + RFV Modules	KNN Imputer (n_neighbors = 5)	0.6134
	Categorical Features + Quantitative Features with Binned Groups + RFV Codes + RFV Modules + Topic Features	***KNN Imputer (n_neighbors = 5)***	***0.6141***
	Categorical Features + Original Quantitative Features + RFV Codes + RFV Modules + Topic Features	KNN Imputer (n_neighbors = 5)	0.6109
Random Forest Classifier	Categorical Features + Quantitative Features with Binned Groups + RFV Codes + RFV Modules	-	0.6138
	Categorical Features + Quantitative Features with Binned Groups + RFV Codes + RFV Modules + Topic Features	-	0.6068
	Categorical Features + Original Quantitative Features + RFV Codes + RFV Modules + Topic Features	-	0.6127
Histogram-Based Gradient Boosting Classifier	***Categorical Features + Original Quantitative Features + RFV Codes + RFV Modules + Topic Features***	-	***0.6430***

Multi-Layer Perceptron Classifier	Categorical Features + Original Quantitative Features + RFV Codes + RFV Modules + Topic Features	KNN Imputer (n_neighbors = 5)	0.5562

Table 5: Performance comparison of models

D. Hyper parameter tuning with HalvingGridSearchCV

1. HalvingGridSearchCV streamlines hyperparameter optimization by halving the candidate configurations each iteration, aiding in efficient identification of the best model parameters (Figure 6). The term 'n_candidates' refers to the number of parameter combinations tested with a given sample size ('n_samples').

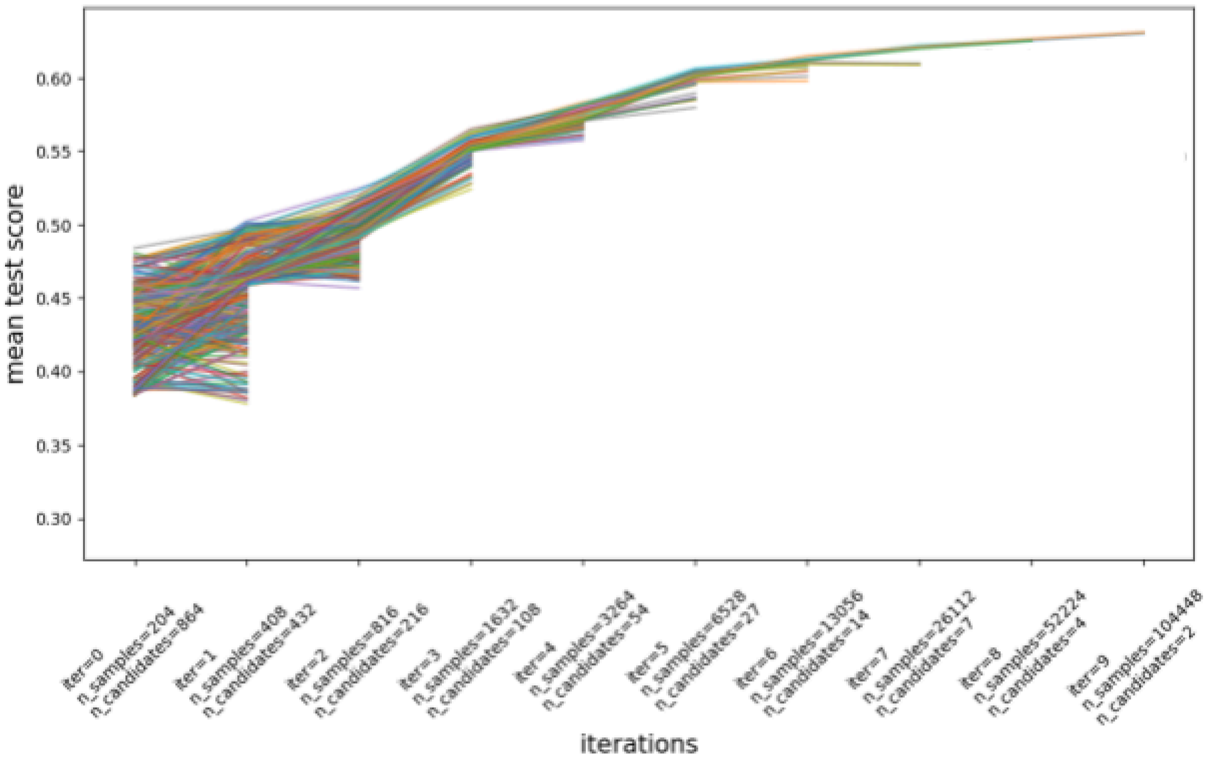


Figure 6: Iterative resource allocation process of Random Forest Classifier

2. The training dataset and validation dataset are merged for HalvingGridSearchCV (Table 6). Subsequently, the models with the optimal parameters are evaluated against the test dataset. The Random Forest Classifier obtained the best score of 0.6135 (Weighted F1) on the test set.

Model	Best Parameters	Test Score (Weighted F1)
Logistic Regression Classifier	classifier__C: 0.1 pca__n_components: 0.95 knn__n_neighbors: 10	0.5877
Random Forest Classifier (see Appendix B for more information)	classifier__bootstrap: True classifier__criterion: gini classifier__max_depth: None classifier__max_features: log2 classifier__min_sample_leaf: 1 classifier__min_sample_split: 10 classifier__n_estimators: 1000 pca__n_components: None	**0.6135**
Histogram-Based Gradient Boosting Classifier (see Appendix C for more information)	classifier__l2_regularization: 0.01 classifier__learning_rate: 0.1 classifier__max_depth: 3 classifier__max_iter: 100 classifier__max_leaf_nodes: None classifier__min_samples_leaf: 20 pca__n_components: None	0.5962

Table 6: Best parameters of each models

E. Final model – Random Forest Classifier

Upon comprehensive evaluation, encompassing factors such as performance (weighted F1-score), robustness against overfitting (discrepancy between mean cross-validation score, the test score, and the final evaluation score), and interpretability, we have selected the Random Forest Classifier as our final model. The best model was assessed against the final evaluation dataset, achieving a weighted F1-score of 0.6344, demonstrating satisfactory generalization to unseen data.

The streamlined Random Forest pipeline was shown in Figure 7. A ColumnTransformer processes data types differently: binary data passes through, categorical data undergoes one-hot encoding, numerical data is standardized, and PCA is optional before Random Forest classification. Figure 8 showcases the confusion matrix for the final model, presenting a detailed breakdown of the model's predictive accuracy across various medical conditions. Each cell in the matrix indicates the percentage of predictions for a given true label (vertical axis) against the predicted label (horizontal axis), with darker shades representing higher percentages. The diagonal cells, highlighted by their darker color, reflect the proportions of correct predictions, while the off-diagonal cells shed light on misclassifications, offering a transparent view of the model's strengths and areas for improvement in diagnostic categorization.

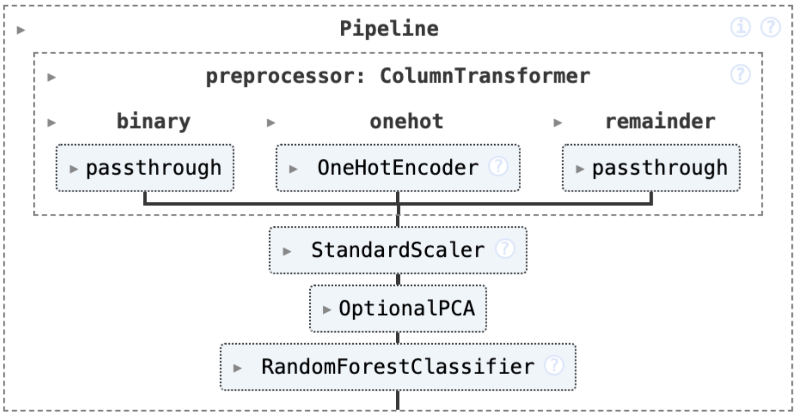


Figure 7: Optimized Data Processing Pipeline for Random Forest Classification

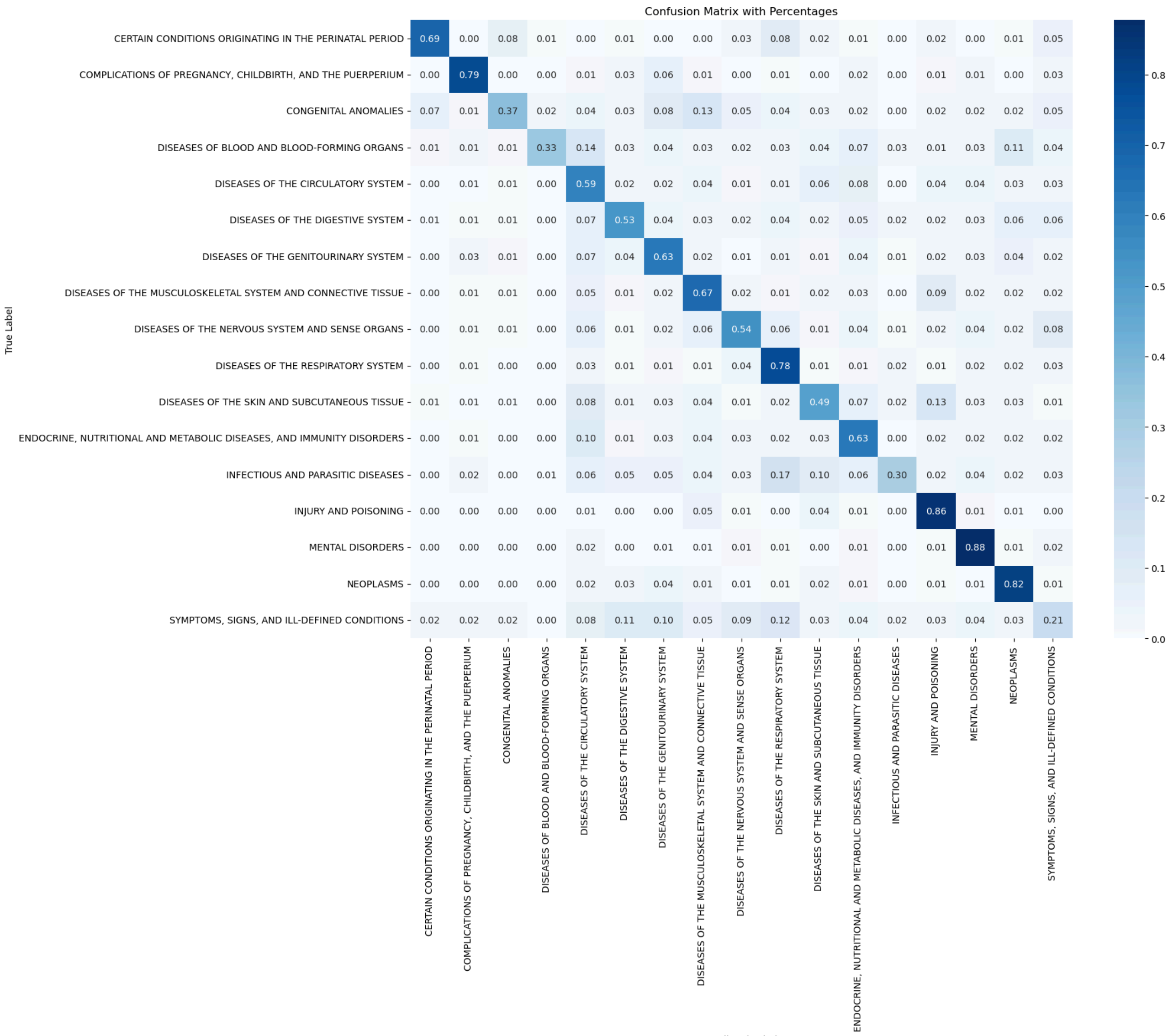


Figure 8: Confusion Matrix of the Optimal Random Forest Model

Subsequently, we concatenate the training, validation, test and final evaluation datasets and refit the final model for potential prediction simulation with user-reported data. Figure 9 delineates the 20 most influential features that contribute to the predictive prowess of our final model, underscoring the variables that hold the most weight. Attributes such as 'AGE', 'WTLB' (weight in pounds), and various topics—derived from NLP analysis—stand out as significant, with 'AGE' being the paramount predictor. The prominence of these features highlights the intricate interplay between demographic characteristics, vital signs, and thematic elements of patient data in our final model.

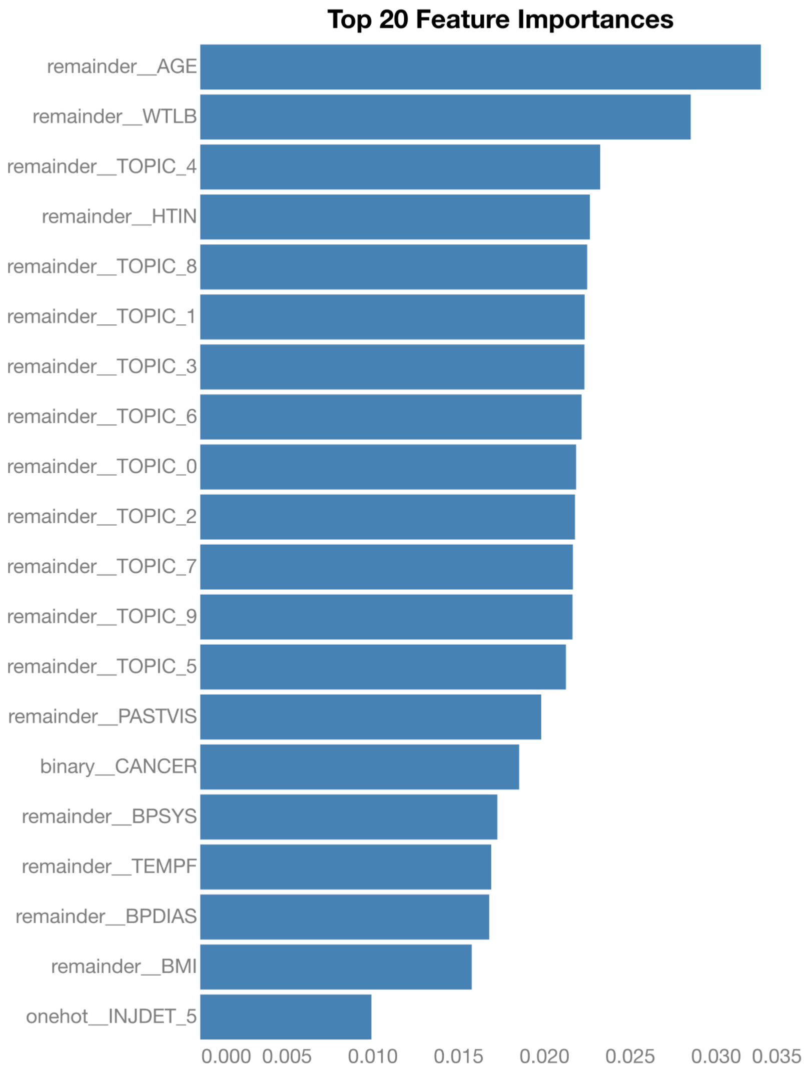


Figure 9: Unveiling the Top 20 Features Influencing Outpatient Diagnoses

Discussion

The insights from our statistical analysis testing underscore the potential for machine learning in enhancing patient care through predictive modeling. Recognizing the effectiveness of models like Random Forest and Gradient Boosting observed in our preliminary tests, our upcoming direction should focus on integrating these models to craft more sophisticated predictive tools. Improving data collection and applying advanced imputation methods will also enhance model accuracy.

Our analysis culminated in a Random Forest model, which guided the creation of topic-specific word clouds to visualize outpatient data themes. We proposed a health self-check stimulator, which aims to forecast possible healthcare services for patients, as depicted in Figure 10. Due to time constraints, only a conceptual design is presented, anticipating future expansion.

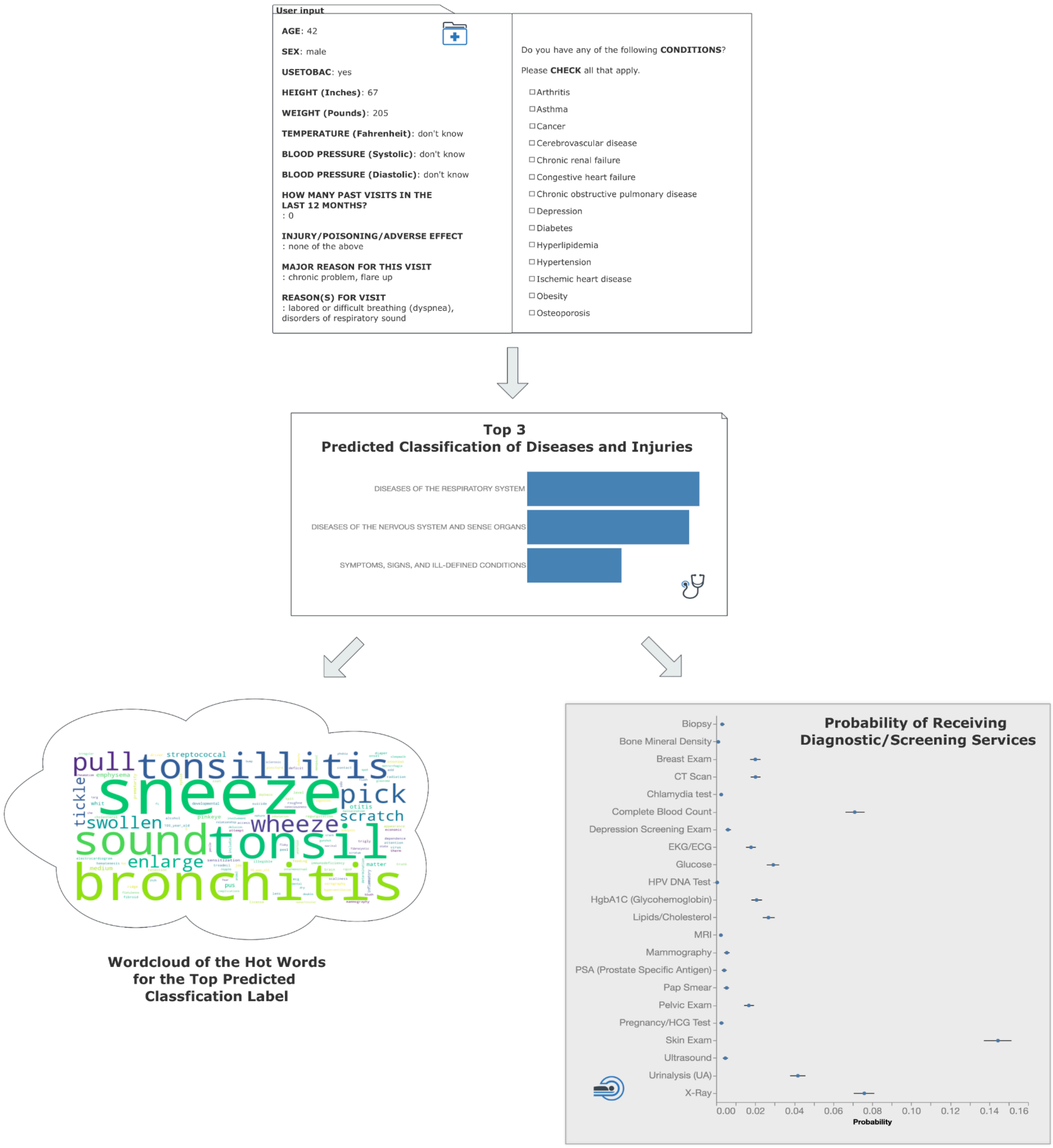


Figure 10: Mock-up health self-check stimulation

Yet, our study is not without its constraints. As indicated in the provided text, data comparability challenges, particularly due to changes in survey questions and the annual revisions to the ICD-9, underscore the limitations in our analysis. Additionally, issues such as data imbalance, dimensionality, and the need for up-to-date information have curtailed the full realization of our model's capabilities. The analysis of our project is also limited by the high percentage of missing data on certain variables. For example, in 2007, notable gaps in NAMCS data were present, with Ethnicity missing 34.7% and Race missing 31.5% of the information, even after considering both imputed and un-imputed data. Thus, we are unable to account for the potential influence of Ethnicity, Race, or other variables on health conditions and associated medical treatments.

Looking to the future, several enhancements stand poised to refine our work. These include rectifying the RFV text for better clarity, employing sentence embeddings to enrich feature extraction, migration to the most current coding system, the International Classification of Diseases, Tenth Revision (ICD-10), and potentially shifting the target outcome to predicting the most pertinent department for the patient's condition. We should also notice the minority representation within the data, thereby creating a model that is both inclusive and reflective of diverse patient groups. An online updating mechanism would also provide the chances for the model to evolve with new data, ensuring its recommendations remain as accurate and relevant as possible.

Ethical considerations are paramount in the use of machine learning models within healthcare contexts. One critical aspect is the potential for model bias, which could skew predictions and inadvertently lead to unequal healthcare services. It is crucial to ensure that the models do not propagate or exaggerate existing disparities, especially for underrepresented groups. Also, we should be careful of over-interpreting results; the outputs of our analysis are intended to augment, not replace, the expertise of medical professionals. They should be used as an informational tool rather than a definitive source of medical advice. This distinction is vital to ensure that users of the system understand the limitations of the model and the importance of seeking professional medical evaluations, thereby avoiding overreliance on automated predictions which may not account for the full complexity of individual health profiles.

Conclusion

Our classification approach on NHAMCS outpatient data reveals key diagnostic pathways and supports, potentially improving patient understanding before medical consultations. Despite time and data limitations, our results endorse the application of machine learning in healthcare, setting a foundation for a conceptual interactive self-check platform.

Statement of Work

Topic and literature searching and discussion: Jingyi LAI

Statistics analysis: Jeffery JIN

Feature engineering: Xuan TAO, Jingyi LAI

Machine learning modeling: Xuan TAO

Poster & Report writing: Jingyi LAI, Xuan TAO, Jeffery JIN

Reference

Anglada-Martínez, H., Martin-Conde, M., Rovira-Illamola, M. et al. Feasibility and Preliminary Outcomes of a Web and Smartphone–Based Medication Self-Management Platform for Chronically Ill Patients. J Med Syst 40, 99 (2016). https://doi.org/10.1007/s10916-016-0456-y

Asao K, McEwen LN, Lee JM, Herman WH. Ascertainment of outpatient visits by patients with diabetes: The National Ambulatory Medical Care Survey (NAMCS) and the National Hospital Ambulatory Medical Care Survey (NHAMCS). Journal of Diabetes and its Complications. 2015;29(5):650-658. https://doi.org/10.1016/j.jdiacomp.2015.03.019

Czakon, J. (2023, September 5). F1 score vs ROC AUC vs Accuracy Vs PR AUC: Which evaluation metric should you choose?. neptune.ai. https://neptune.ai/blog/f1-score-accuracy-roc-auc-pr-auc

Hing E, Middleton K. National Hospital Ambulatory Medical Care Survey: 2001 outpatient department summary. Adv Data. 2003 Aug 5;(338):1-26. PMID: 12918175.

M.A. Gupta, A.K. Gupta, S.J. Chen, A.M. Johnson, Comorbidity of rosacea and depression: an analysis of the National Ambulatory Medical Care Survey and National Hospital Ambulatory Care Survey—Outpatient Department data collected by the U.S. National Center for Health Statistics from 1995 to 2002, British Journal of Dermatology, Volume 153, Issue 6, 1 December 2005, Pages 1176–1181, https://doi.org/10.1111/j.1365-2133.2005.06895.x

Appendix

A Github repository and data access
B Results of Logistic Regression classifier
C Histogram-Based Gradient Boosting Classifier
D Word cloud of the final best model(Random forest model)

Appendix A Github repository and data access

Github repository
https://github.com/sean7x/pathclarity

The ASCII OPD data files can be downloaded from the following hyperlink:
https://ftp.cdc.gov/pub/Health_Statistics/NCHS/Datasets/NHAMCS/

The accompanying SPSS Documentation and code, which can serve as helper files for parsing the ASCII data into Pandas DataFrames, can be downloaded from the following:
https://ftp.cdc.gov/pub/Health_Statistics/NCHS/dataset_documentation/nhamcs/spss/

Appendix B Results of logistic regression classifier

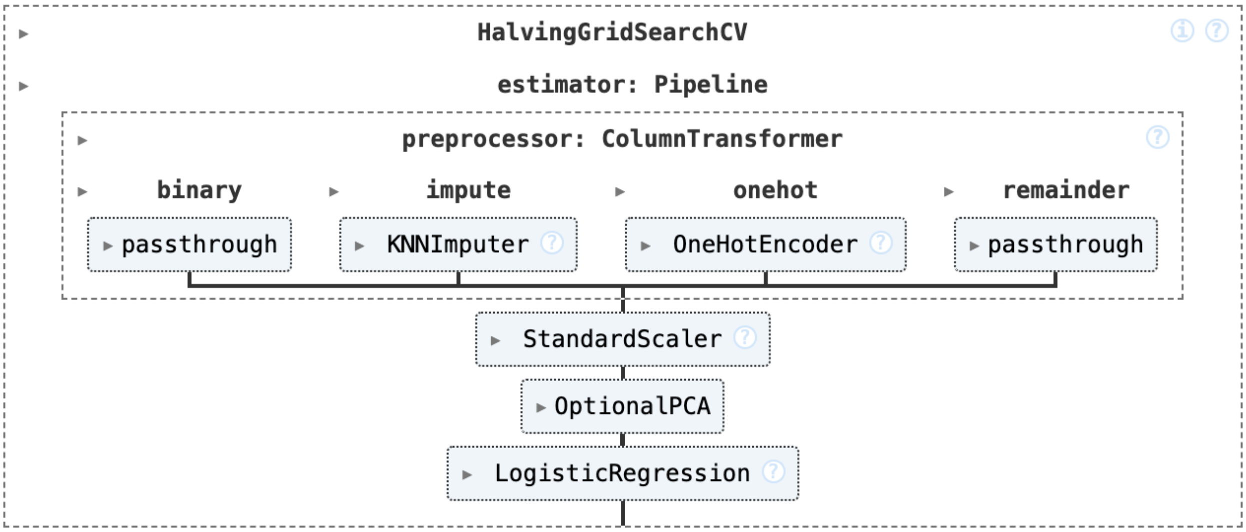


Figure B.1: Pipeline of Logistic Regression Classifier

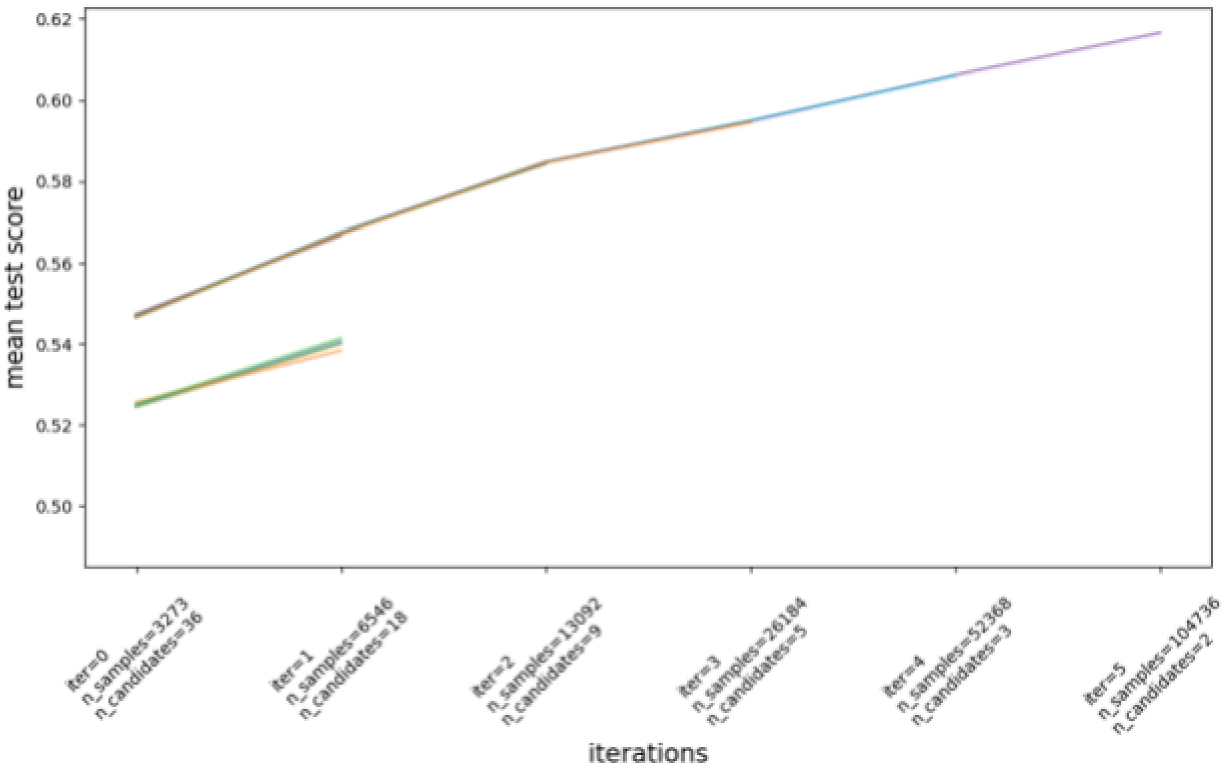


Figure B.2: Iterative resource allocation process of Logistic Regression Classifier

Appendix C Results of Histogram-Based Gradient Boosting Classifier

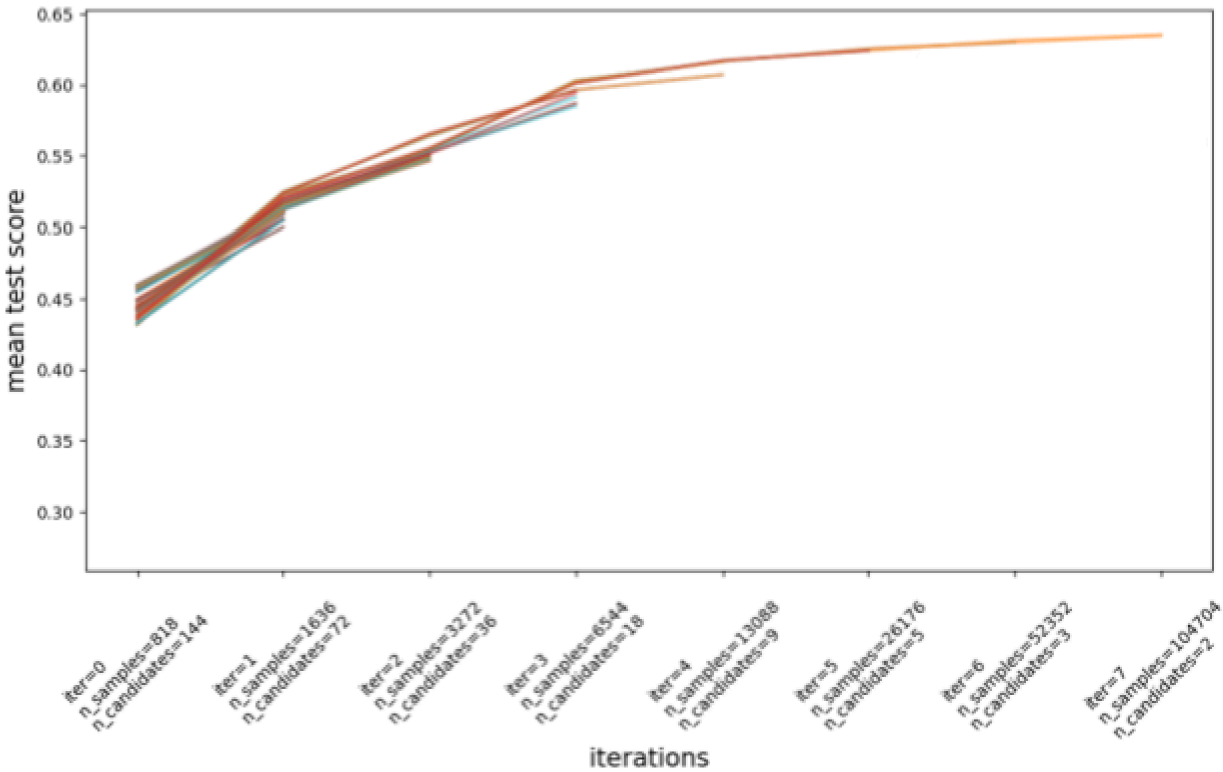
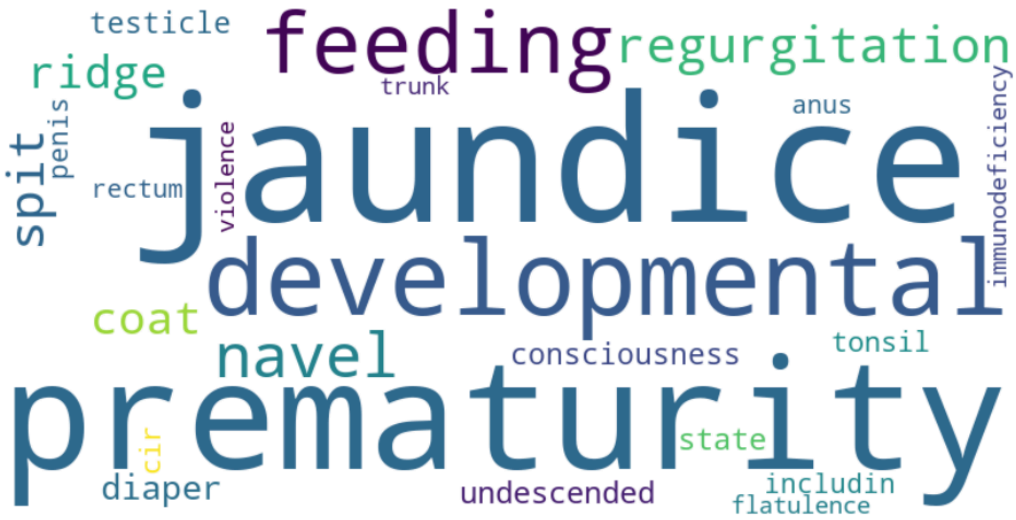
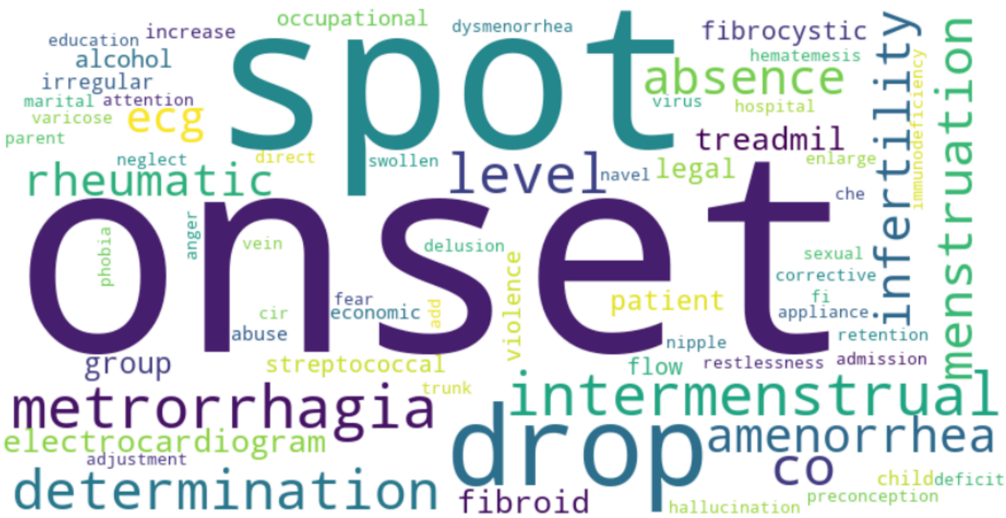


Figure C: line graph about the reduction in loss over iteration for the gradient boosting model

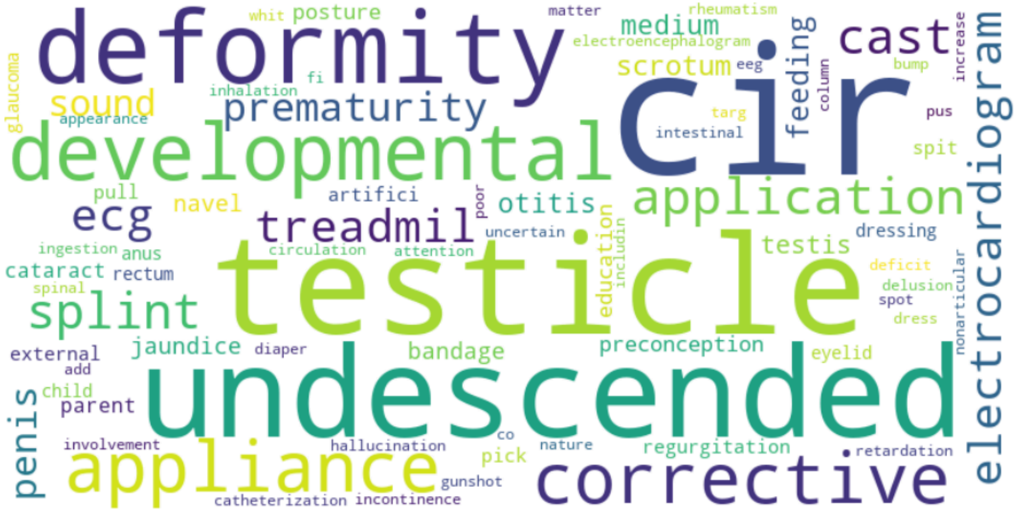
Appendix D Seventeen Comprehensive Word Cloud



1. Word Cloud for Certain conditions originating in the perinatal period

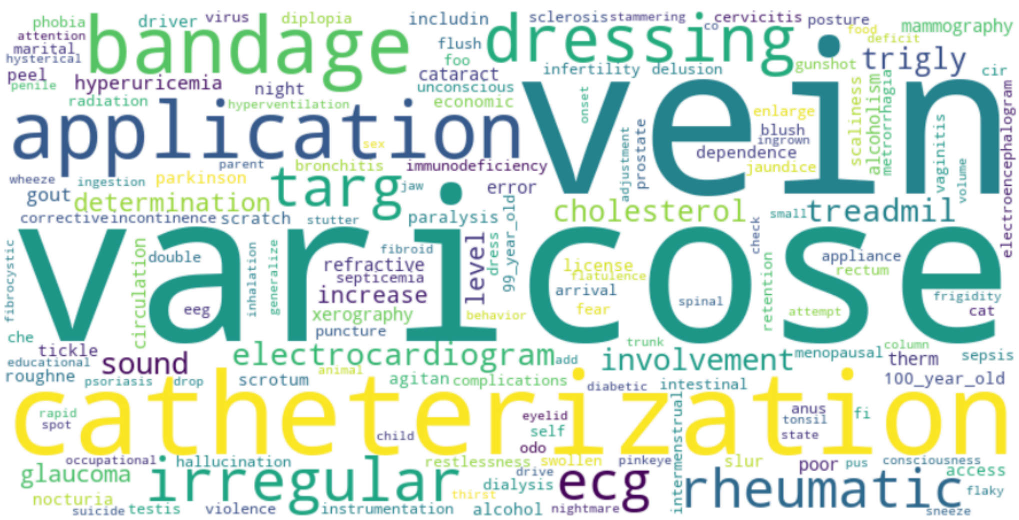


2. Word Cloud for Complications of pregnancy, childbirth, and the puerperium



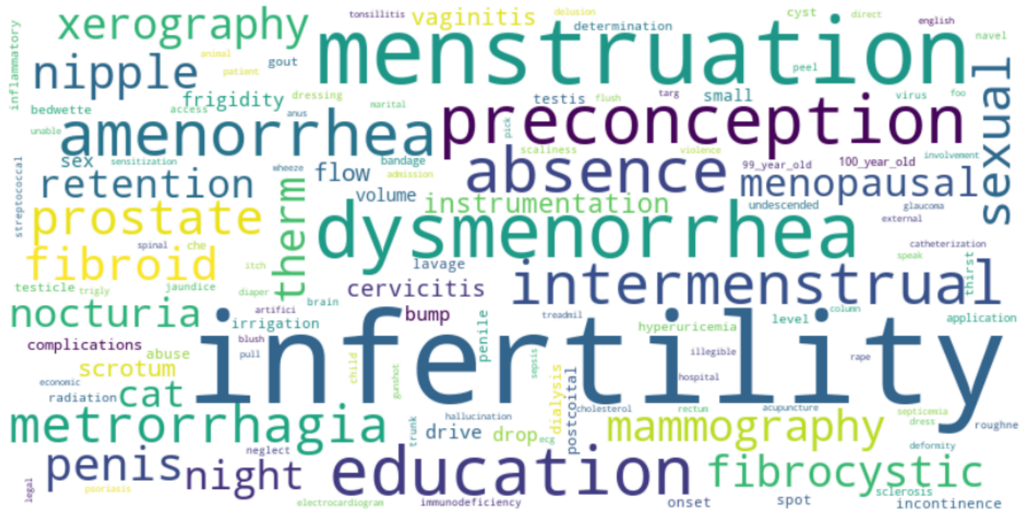
3. Word Cloud for Congenital anomalies

4. Word Cloud for Diseases of blood and blood-forming organs

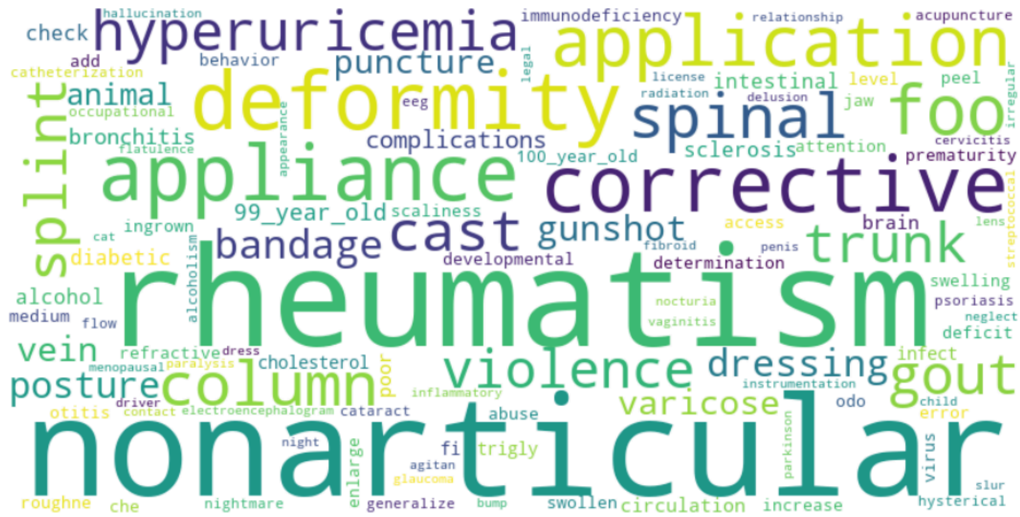


5. Word Cloud for Diseases of the circulatory system

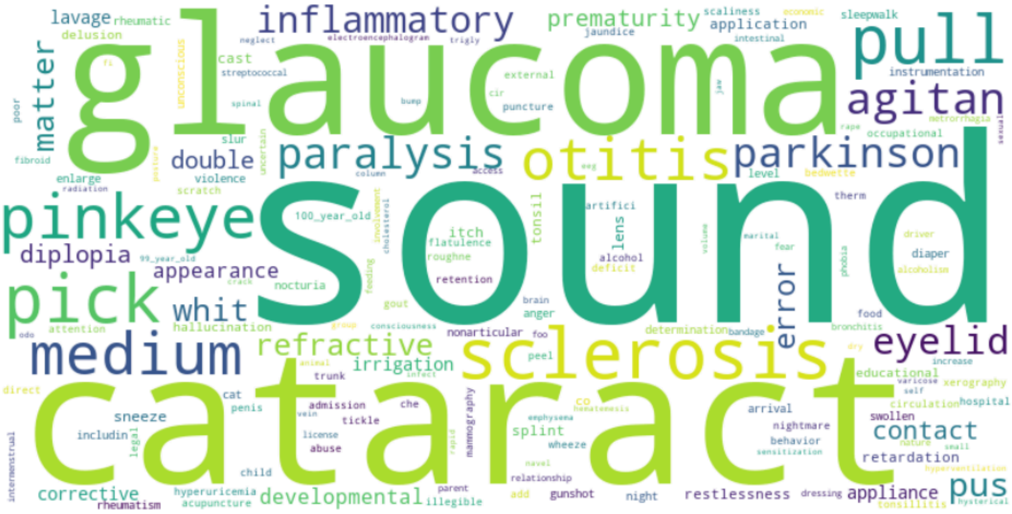
6. Word Cloud for Diseases of the digestive system



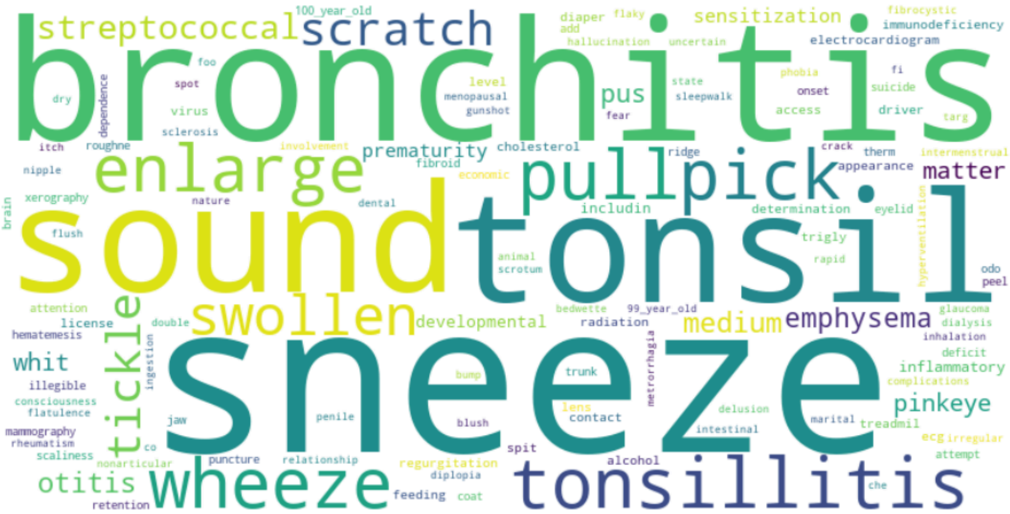
7. Word Cloud for Diseases of the genitourinary system



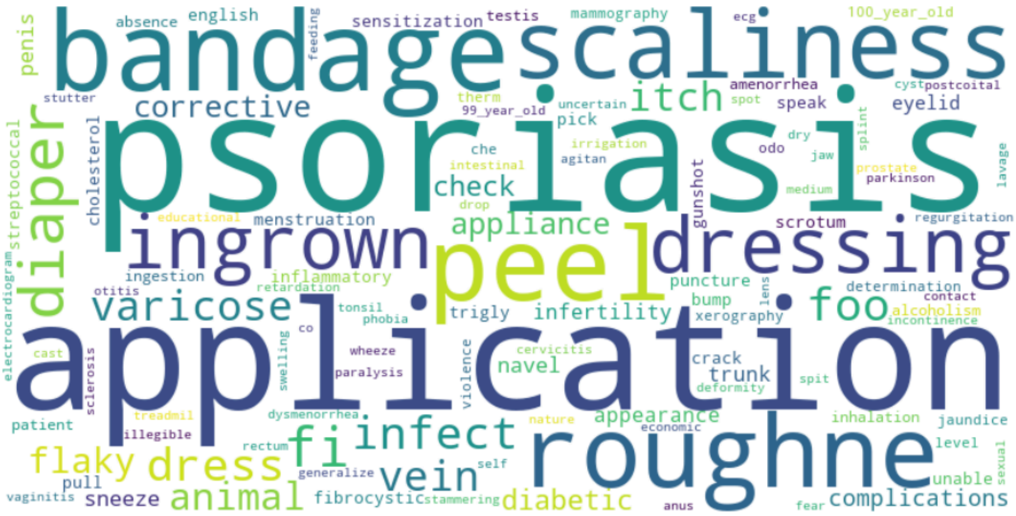
8. Word Cloud for Diseases of the musculoskeletal system and connective tissue



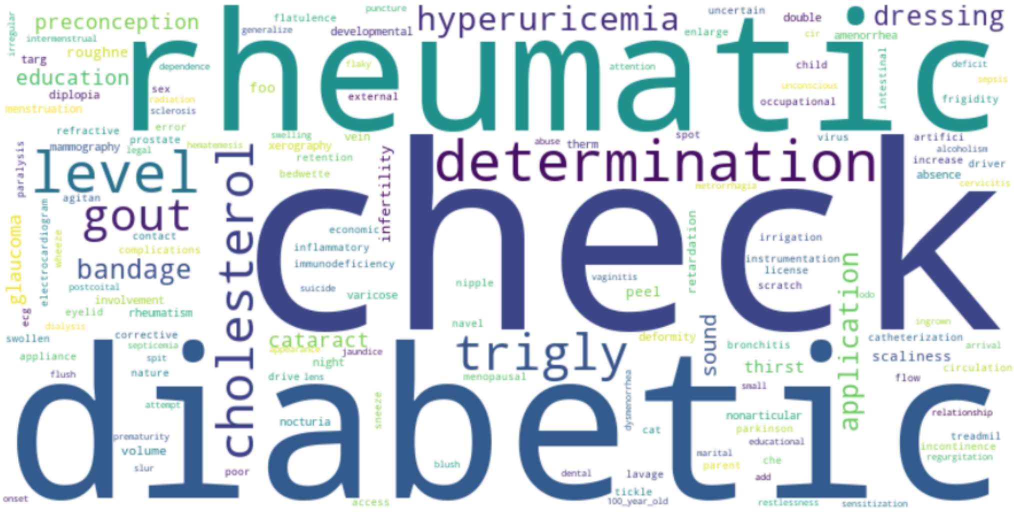
9. Word Cloud for Diseases of the nervous system and sense organs



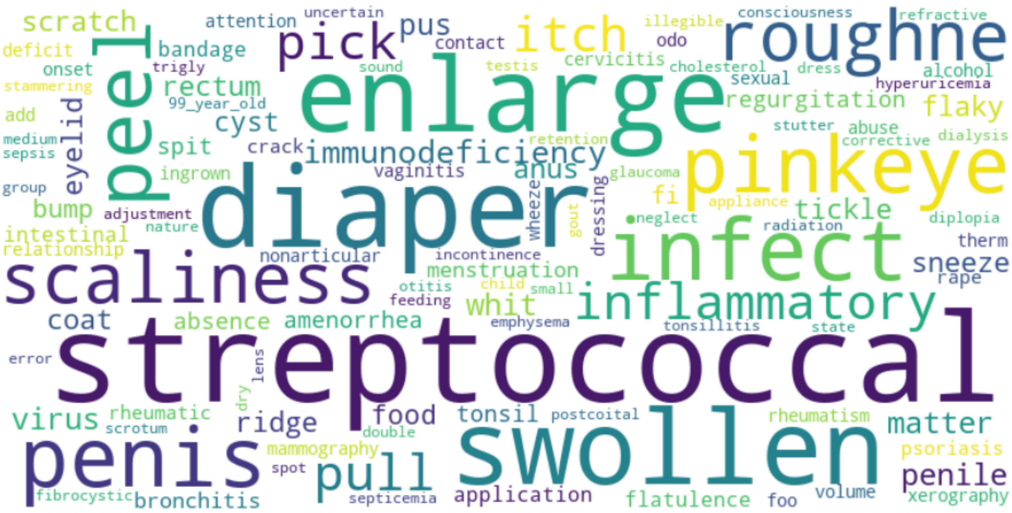
10. Word Cloud for Diseases of respiratory system



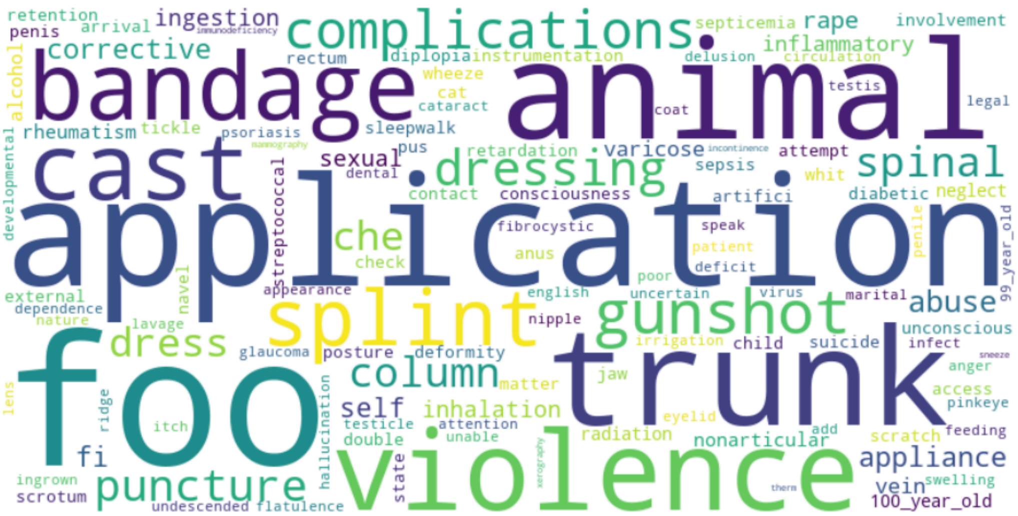
11. Word Cloud for Diseases of the skin and subcutaneous tissue



12. Word Cloud for Endocrine, nutritional and metabolic diseases, and immunity disorders

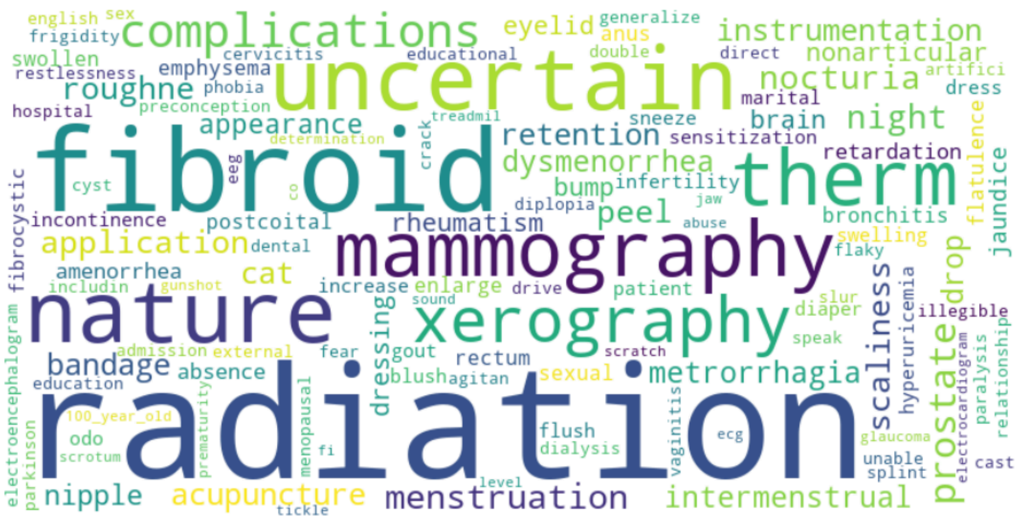


13. Word Cloud for infectious and parasitic diseases

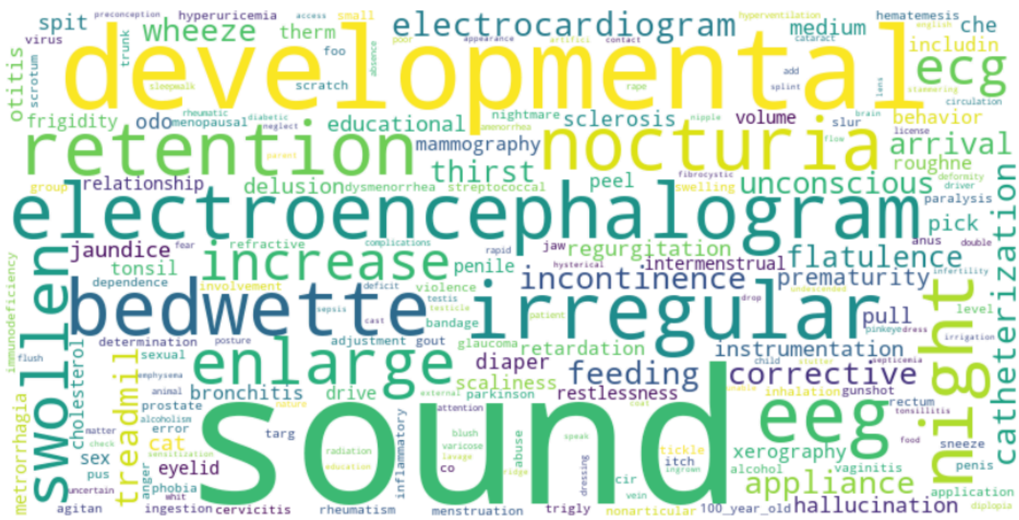


14. Word Cloud for injury and poisoning

15. Word Cloud for mental disorders



16. Word Cloud for Neoplasms



17. Word Cloud for Symptoms, signs, and ill-defined conditions