Methodology and Analysis Detailed Summary

Stage 1 – Data Familiarization and Conceptual Grouping

The data was categorized clearly into three main groups based on conceptual usage:

1. **Group 1 - Indicators and Retrospective Markers**:
    - Primarily composed of match_* variables such as `match_pdf_after`, `match_diag_141`, and `match_aspirin_after`, indicating specific medical diagnoses or treatments retrospectively.
    - Includes the critical target variable (`Y`) indicating final diagnosis status.
2. **Group 2 - Clinical Diagnosis Variables**:
    - Includes variables such as `preeclampsia_sum`, `pregnancy_hypertension_sum`, and `labs_sum`, each representing the cumulative count of clinical diagnoses or related measures.
    - A patient may have multiple clinical diagnoses.
3. **Group 3 - Auxiliary Variables for Modeling**:
    - Incorporates demographic information, laboratory test results, blood pressure measurements, smoking history, medical texts, and prior diagnostic codes.
    - Variables were organized using PREFIX naming conventions for ease of handling and clarity.

Stage 2 – Exploratory Data Analysis (EDA) and Initial Processing

The dataset exhibited significant class imbalance (~432 diagnosed positive cases vs. ~9500 healthy negative cases).

Detailed handling for each group:

- **Group 1**:
    - Majority had one diagnostic indicator per patient.
    - The `match_rasham_after` variable alone was not highly informative, as its positive occurrences were always accompanied by positive indicators in other specific columns (`match_aspirin_after` or `match_pdf_after`).
    - The column `match_measure_after` indicated excluded cases and was entirely empty, thus removed from further analysis.
- **Group 2**:
    - Empty columns (representing pre-excluded patient records) were removed.
    - Most patients had one diagnosis, although some had multiple diagnoses.
- **Group 3**:
    - The `Init_date` column was a sequential number providing no predictive value, hence removed.
    - Demographic variables appeared stable and reliable, with only minor missing data (9 rows). These missing values were intentionally retained, relying on the tree-based model's robustness to handle missing data effectively.

- **Smoking Variables**: Showed extensive missingness. Despite their potential clinical relevance, their predictive value in this dataset appeared limited. Unrealistic entries (e.g., 120 years of smoking) were corrected to NULL.
- **Labs**: Due to numerous lab variables, general visual inspections were conducted. Most lab variables appeared complete and free of problematic anomalies. The variable `lab_papp_a_abs_last_value` underwent a log transformation (`log1p`) due to extreme values. Highly correlated variables were retained for later selection via ElasticNet regularization.
- **MEASURE Variables**: appeared normal with expected correlations.
- **Num of diagnosis columns**: Sparse columns with very few patients (1-10) were retained for regularization to determine importance. An outlier in `24_diag_80_num_of_diag` (112 diagnoses) was identified as erroneous and set to NULL. The considered aggregate `total_diag_count` was excluded due to extreme correlations.
- **Days since last diagnosis**: No issues or extreme values identified.

Additional Clinical Insights:

- Partial overlap between gestational and chronic hypertension was identified (31.5% overlap), indicating potential co-risk.
- Preeclampsia was the most frequent diagnosis, whereas eclampsia was rare. Chronic hypertension and pregnancy-induced hypertension had moderate frequencies.
- Diagnosed women were slightly older on average compared to healthy ones, though age distributions significantly overlapped, suggesting only a moderate relationship.
- Smoking status showed minimal correlation with the diagnosis.
- Blood pressure (both systolic and diastolic) was notably higher among positively diagnosed patients.

Literature Review Correlation:

A correlation study based on clinical literature indicated:

- Strong and consistent predictive power of maximal systolic and diastolic blood pressures.
- Moderate correlation with weight.
- Weak correlation with age, despite being reported relevant in literature.
- Laboratory measures (platelets, urinary protein, hemoglobin, hematocrit, PAPP-A, MPV) showed weaker correlations, suggesting potential nonlinear relationships requiring advanced modeling approaches.

NLP Feature Engineering:

- A new variable (`clinical_length`) was added to represent the length of medical texts.
- Data was split into Train/Test subsets early, using stratified sampling to ensure balanced distribution of diagnostic subtypes and treatment indicators. Rare groups were merged to maintain statistical stability and prevent information leakage.
- **Embeddings**: Encoded only the last diagnostic paragraph, using a fallback to the full text when necessary, leveraging the multilingual-e5-base model trained on medical texts.
- **TF-IDF Features**: Generated exclusively on the Train set. Words with highest Mutual Information relative to diagnosis status were selected to create binary indicators. No exposure or adjustment based on the Test set was performed.

Regularization and Feature Selection:

The data was segregated into three distinct datasets for systematic feature selection:

- **Embeddings set**: Selected via LASSO without standardization due to uniform scale.
- **Words set**: Binary TF-IDF features selected via LASSO, identically applied to Test.
- **Clinical data**: Implement standardization with StandardScaler, and selected via ElasticNet, balancing feature sparsity and correlation.
- All selection processes were strictly confined to the Train dataset to prevent leakage.

Modeling and Evaluation:

- Model training with LightGBM, addressing class imbalance using class weights ({0:1, 1:2}).
- Robust evaluation using Stratified K-Fold Cross-Validation exclusively on Train data.
- Final evaluation conducted on Test set, reporting precision, recall, and F1-score at threshold 0.5.

Clinical-Business Interpretation by Percentiles ("Top X%" based on probability ranking, not fixed threshold):

- **Top 1%** identifies most positive cases (85%) with very few referrals (only 30 women), but at reduced precision—thus, more false positives.
- **Top 5%** maintains a balanced approach, offering high precision (88%) and good recall (78%), suitable for moderately constrained budgets.
- **Top 10%** provides the highest precision (93%) but with slightly lower recall, ideal for conservative scenarios where referral reliability is critical.

Recommended threshold depends on cost-benefit considerations:

- For tight budget constraints, start with the **Top 5%**, offering an optimal balance between accuracy and coverage.
- For systems with higher budget flexibility, consider the **Top 10%** to maximize referral accuracy, or the **Top 1%** if maximizing detection outweighs the higher false-positive costs.

Feature Importance:

Analysis with LightGBM's Gain and SHAP values consistently highlighted key TF-IDF words, selected embeddings, and diastolic blood pressure as critical predictors.

Future Improvements

- **Severity Modeling:** Shift from binary classification to multi-class or ordinal modeling, reflecting clinical severity and optimizing resource allocation.

- **False Negative Analysis:** Conduct targeted analyses to identify recurring patterns among missed cases, enabling model refinements.

- **Enhanced NLP:** Fine-tune embeddings specifically for medical text, improving textual feature extraction.

- **LLM Integration:** Develop a dedicated LLM model using clinical texts for direct classification, with additional clinical features as support.

- **Longitudinal Data Integration:** Encode multiple patient visits separately or sequentially to capture temporal context.

- **Data Completion (e.g., Smoking):** Obtain accurate smoking data from reliable sources to enhance predictive power.

- **Temporal Trends:** Incorporate time-dependent metrics (e.g., weight gain, blood pressure variability) into the model.

- **Dynamic Thresholds:** Implement budget-aware referral thresholds dynamically optimized per operational cycle.