Question Bank for Experiment 1

**Question based on paper**

1. **Problem Identification**: What specific challenges in data cleaning for Electronic Health Records (EHRs) does this paper address?
2. **Methodology**: How does the proposed method incorporate clinical knowledge to automate data cleaning?

What are the key steps in the method?

1. **Algorithms**: Can you explain the algorithms or models used in this data cleaning method?
2. **Evaluation**: How was the effectiveness of the data cleaning method evaluated? What metrics were used?
3. **Clinical Impact**: How does this automated data cleaning improve the use of EHRs in clinical settings?
4. **Comparison**: How does this method compare to other existing data cleaning techniques for EHRs?
5. **Limitations**: What are the limitations of this approach as discussed in the paper?
6. **Future Work**: What future improvements or expansions are suggested for this research?

**Question Based on Experiment**

* 1. What is the primary goal of cleaning, integrating, and transforming EHRs in your experiment?
  2. Which types of healthcare data were used in your experiment? How did you handle different formats?
  3. What methods did you use to handle missing data, duplicates, and outliers in the EHRs?
  4. How did you approach schema matching and entity resolution for data from different sources?
  5. What features did you create from the cleaned and integrated EHR data, and how did they improve analysis?
  6. Did you apply any Natural Language Processing (NLP) techniques for clinical notes? How effective were they?
  7. Can you explain the process of time-series transformation in your experiment? How did you aggregate time-series data?
  8. How did you validate the quality of your integrated data? What metrics did you use to evaluate its consistency?
  9. What challenges did you face during the integration of multiple EHR systems? How did you overcome them?
  10. How did you ensure the de-identification of patient data during cleaning and transformation?

**Questions Based on the Paper:**

1. **Problem Identification:**  
   The paper addresses challenges in data cleaning for Electronic Health Records (EHRs) related to variable-specific data inconsistencies. Issues include errors in values (e.g., extra or missing zeros), out-of-range values, and inconsistent or misspelled measurement units. The challenges stem from EHRs being derived from various sources like general practices, hospitals, and laboratories, leading to a lack of uniformity in data formatting and variable-specific knowledge requirements for cleaning.
2. **Methodology:**  
   The proposed method integrates clinical knowledge via a Clinical Knowledge Database (CKD), which stores variable-specific information such as normal ranges, unit conversion formulas, and plausible value limits. The approach involves fuzzy search to correct unit spelling errors, conversion of units to a standard format, and outlier detection based on clinical thresholds. The key steps are:
   * Extracting data for each clinical variable.
   * Preprocessing the data to remove unexpected characters and convert values to numeric format.
   * Applying fuzzy search for unit correction and standardization.
   * Converting units based on the CKD.
   * Detecting and handling outliers based on normal and extreme value ranges.
3. **Algorithms:**  
   The algorithms used include fuzzy search (based on Levenshtein distance) to identify and correct misspelled units, unit conversion using predefined conversion rates, and outlier detection through range checks. The fuzzy search algorithm calculates similarity scores between raw and known units, while outlier detection adjusts values to fall within the normal range if possible or replaces extreme values with missing data (NA).
4. **Evaluation:**  
   The effectiveness was evaluated by measuring the completeness (percentage of missing values) and correctness (percentage of values within normal ranges) of clinical variables before and after cleaning. The data quality improvement was indicated by a higher percentage of values falling within the normal range post-cleaning.
5. **Clinical Impact:**  
   Automated data cleaning improves EHR usability by ensuring data consistency and accuracy, making it easier for non-technical users to analyze clinical data. The method enables quicker and more reliable preparation of data for clinical research and decision-making.
6. **Comparison:**  
   Unlike other data cleaning methods that rely on statistical distributions, this approach incorporates clinical knowledge for variable-specific data handling. It differs from previous methods that required manual intervention or were limited to specific data types (e.g., anthropometric data).
7. **Limitations:**  
   The limitations include the inability to incorporate longitudinal patient data to improve data consistency and the risk of a slight decrease in data completeness due to the replacement of outliers with NA values.
8. **Future Work:**  
   Suggested future improvements involve expanding the CKD with more clinical rules, incorporating longitudinal data analysis, and adding more complex consistency checks (e.g., rules for related variables).

**Questions Based on the Experiment:**

1. **Primary Goal:**  
   The goal of cleaning, integrating, and transforming EHRs is to standardize the data for accurate analysis and decision-making, ensuring that the records are consistent, complete, and clinically meaningful.
2. **Healthcare Data Types:**  
   The experiment involved different healthcare data types, including laboratory test results, measurements, and clinical records. Handling different formats involved converting various units to a standard form and dealing with inconsistent data entries.
3. **Handling Missing Data, Duplicates, and Outliers:**  
   Methods included replacing outliers with NA based on clinically defined ranges, identifying and removing duplicates, and marking invalid entries as missing to ensure data quality.
4. **Schema Matching and Entity Resolution:**  
   Schema matching involved aligning variable names and units across different data sources using CKD rules, while entity resolution aimed to integrate records from various systems by matching patient and test identifiers.
5. **Feature Creation:**  
   Features created included normalized test results and derived metrics (e.g., calculated risk scores), which enhanced the ability to analyze clinical outcomes.
6. **NLP Techniques for Clinical Notes:**  
   The experiment did not apply specific NLP techniques to clinical notes, focusing instead on structured data cleaning.
7. **Time-Series Transformation:**  
   Time-series data was aggregated by normalizing measurement intervals and summarizing values using statistical measures (e.g., mean, median).
8. **Data Validation:**  
   The quality of integrated data was validated using metrics such as completeness, correctness, and the percentage of values within normal ranges.
9. **Challenges in Integrating Multiple EHR Systems:**  
   Challenges included inconsistent units, varying data formats, and different coding standards across systems. Solutions involved standardizing units using CKD and converting data to a uniform schema.
10. **De-identification of Patient Data:**  
    De-identification was ensured by removing personally identifiable information and aggregating data where necessary.