Capstone I

Project Report – Assignment 2

Sampling Strategies for CLABSI Estimation

Spring 2024

Team 8

Pavan Kumar Theegela, Jyothika Sakamuri, Taha Mandviwala

BZAN 6360 – Capstone Practicum Project Course I

Nov 11, 2024

Table of Contents

Executive summary

This report explains the process of appropriately choosing a sampling technique which will be used to extract a representative sample drawn from an enormously large dataset which is normally characterized by highly skewed outcomes of interest—in this case, Central Line-Associated Bloodstream Infections, or CLABSI, events.

The following dataset, which serves as the sampling frame for our analysis, features patient identifier, number of patient occurrences, number of CLABSI events, and CLABSI event (T/F) featuring across more than 5000 observations. That was the exploratory analysis of the dataset, which would help in understanding the frequency of patients and CLABSI events, therefore giving insight to evaluate the good sampling plan in extracting 10% of the distribution for future modeling activities. Out of the five probabilistic sampling techniques that had been reflected on, we chose simple random sampling and stratified sampling based on the pros and cons with regard to our dataset characteristics. Considering the skewness and bias mitigation of the dataset, these sampling techniques seem best to derive good samples.

These are the workflows that extract a sample using simple random sampling and stratified random sampling, along with descriptive statistics that compare each sample to the population. Accordingly, further analyses were carried out to compare samples based on key data descriptors to understand possible implications for estimating the occurrences of CLABSI events and representative sampling for modeling the drivers of variation in CLABSI precis.

Overall, this report represents yet another sampling plan to extract a sample from our dataset in order to predict occurrences of CLABSI using the characteristics representative of the population dataset.

Analysis

The sampling dataset provided includes 4 variables: PatientKey, Number of Patient Occurrences, Number of CLABSI Events & CLABSI Event(T/F). The dataset has more than 5000 observations in which 3% of them had suffered a CLABSI event as shown in (Figure 1). The frequency graphs for both CLABSI occurrences and patient occurrences explains that the dataset is right skewed having extreme outliers. To support this observation, we have calculated statistical parameters such as mean, median, skewness and kurtosis of the population. We can see

that there are a few patients with high number of occurrences in patient occurrences and for number of CLABSI events mostly there are 0 to very few occurrences, but there can be seen that some of the patients have high number of CLABSI events

Probabilistic Sampling Techniques

There are a total of five sampling techniques that includes simple random sampling, stratified random sampling, cluster sampling, systematic sampling, and multistage sampling, these come with different approaches for extracting sample based on desired outcome of interest and specific to population. These methods provide efficient and representative sampling by randomly selecting individuals, dividing populations into strata for proportional representation, choosing clusters, systematically sampling at regular intervals, or combining multiple approaches to handle complex scenarios.

Case Sampling Technique Selection

Sampling Technique	Description	Pros	Cons
1. Simple Random Sampling	All the members of the population have an equal opportunity of being chosen for the sample, generally based on random number generation.	- Easy to implement and understand. - Reduces bias, yielding a representative sample.	- Requires an exhaustive list of the population. - Not practical for widespread populations. - Can be time-consuming.
2. Stratified Sampling	The population is divided into strata (sub-groups) based on some characteristic, and samples are randomly drawn from each stratum.	- Ensures representation of all groups, increasing accuracy. - Provides more precise estimates.	- Needs extensive population info to specify strata. - More difficult to arrange and examine.

3. Systematic Sampling	Chooses individuals from a larger population at regular intervals (such as every nth individual).	- Simple and quick to use. - Ensures distribution across the population. - Cheaper than simple random sampling.	- If reordered, results may be biased due to periodicity risk. - Assumes population is evenly spread out.
4. Cluster Sampling	Population is divided into clusters, typically based on geography, and a random sample of clusters is selected; all members of chosen clusters are included.	- Economical for mass populations distributed over a large geographic area. - Convenient for fieldwork.	- Less accurate if clusters are not representative. - Higher sampling error compared to other methods.
5. Multi-Stage Sampling	Involves different sampling methods (e.g., stratified and cluster) applied in multiple stages for a complex population.	- Highly flexible for complex surveys. - Allows sampling at various stages, reducing cost and time.	- Complex to design and analyze. - Higher risk of sampling error at each stage. - Potential for bias if stages are poorly defined.

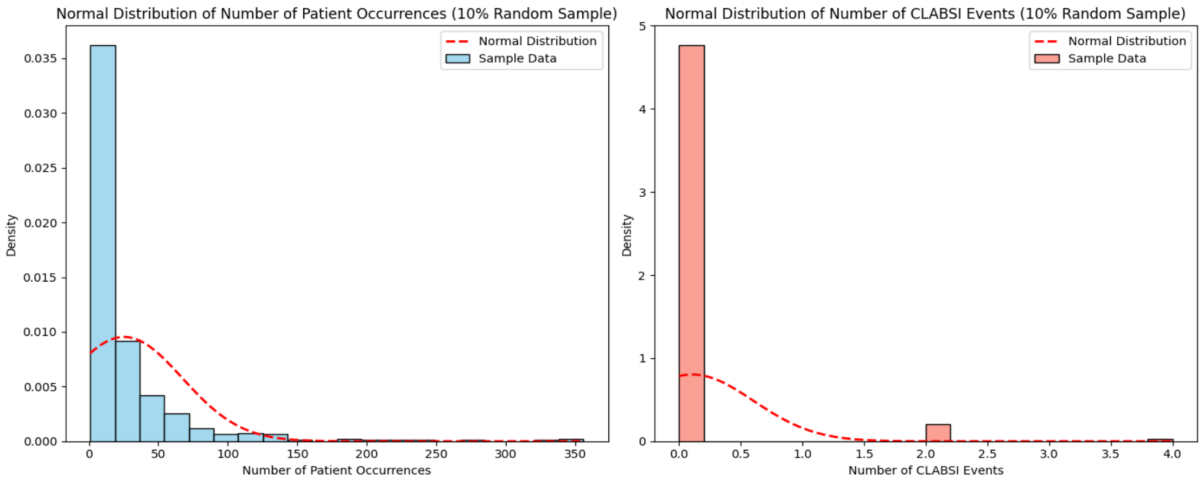
For predicting CLABSI (Central Line-Associated Bloodstream Infections) in a pediatric ICU population, we chose two sampling methods: simple random sampling and stratified random sampling. We opted for these because the dataset is large and has a highly skewed distribution of the outcomes we are interested in. These methods are effective in ensuring the data is representative, reducing bias, and providing reliable estimates, which are crucial for building robust predictive models.

Simple random sampling is easy to implement, while stratified random sampling helps ensure representation by considering key subgroups. However, there are trade-offs. With simple random sampling, there's a risk of bias if the data is highly variable. Stratified sampling requires us to have prior knowledge about the groups (or strata) in the dataset. Despite these challenges, these methods were the best fit for our use case since the dataset naturally divides into clear groups: patients who developed CLABSI versus those who did not. This clear stratification makes it an ideal candidate for stratified sampling.

Simple Random Sampling

Sample Overview: Simple random sampling was implemented by randomly selecting 10% of the patient records from the CLABSI dataset. This process ensures each patient record had an equal chance of being included in the sample, providing an unbiased representation of the entire dataset.

	Patient Occurrences	CLABSI Events
Mean	25.407197	0.106061
Median	11.000000	0.000000
Mode	1.000000	0.000000
Standard Deviation	41.830948	0.496784
Variance	1749.828183	0.246794
Min	1.000000	0.000000
Max	356.000000	4.000000



Summary Statistics:

- **Patient Occurrences**:
  - **Mean**: The mean number of patient occurrences was 25.41, reflecting the potential inclusion of high-value outliers.
  - **Median**: The median value of 11 suggests that while most patient occurrence counts are below this point, some higher values skewed the average.

    - **Standard Deviation and Variance**: A standard deviation of 41.83 and a variance of 1749.83 indicate significant variability, implying a wide range of patient occurrence counts, influenced by the presence of outliers.
- **CLABSI Events**:
    - **Mean**: The mean number of CLABSI events was 0.106, suggesting that rare, high-event cases were present.
    - **Median**: The median of 0 indicates that most records in the sample had no CLABSI events, consistent with the sparse occurrence of such events in the dataset.
    - **Standard Deviation and Variance**: A standard deviation of 0.497 and variance of 0.25 highlight the variability in the sample, which included some higher event counts.

Distribution of Data: The distribution of "Number of Patient Occurrences" and "CLABSI Events" in the simple random sample shows a right-skewed nature:

- **Skewness**: Positive skewness reflects that the distribution had a long tail to the right due to high outliers, especially in patient occurrences.
- **Kurtosis**: Higher kurtosis indicates the presence of extreme values in the sample, pointing to a distribution with heavy tails.

Correlation between Patient Occurrence and CLABSI Events: A correlation analysis revealed that while there was some positive correlation between the number of patient occurrences and CLABSI events, the relationship was not strong enough to indicate a clear dependency. This suggests that while patient occurrences might be related to the frequency of CLABSI events, other factors are likely contributing to the variability.

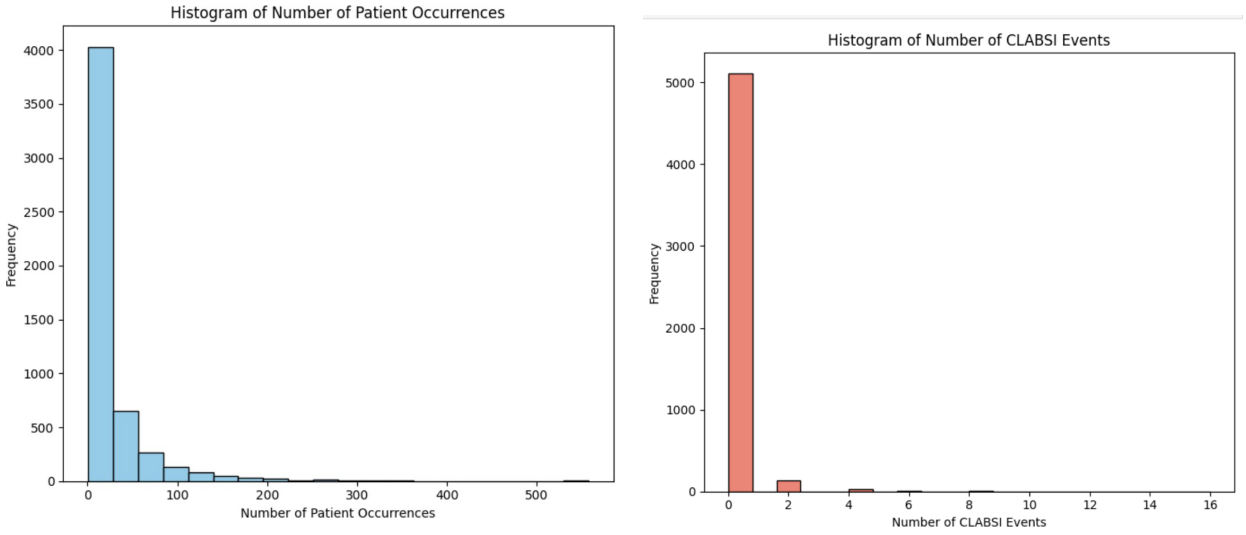
Observations and Conclusions:

- **Observation**: Simple random sampling provides an unbiased snapshot of the entire dataset, capturing high-value outliers that contribute to higher variability and skewness. This can be useful for understanding the impact of outliers on data distribution but may not adequately represent all patient subgroups.
- **Conclusion**: While simple random sampling offers a clear overview of the population, its inclusion of extreme values can influence the analysis, potentially reducing the stability of predictive models. For a more controlled representation, alternative sampling methods may be necessary to complement simple random sampling.

Stratified Sampling

Stratified sampling divides a population into distinct subgroups, or strata, that share similar characteristics. A random sample is then taken from each stratum proportionally, ensuring that each subgroup is represented according to its proportion in the overall population. This method is often chosen over simple random sampling, especially when certain subgroups are more variable or less frequent. Stratified sampling provides a more balanced and representative picture of the population, improving the accuracy and reliability of statistical estimates.

We implemented stratified sampling to draw a 10% sample from each subgroup, specifically focusing on capturing representative information on "Number of Patient Occurrences" and "Number of CLABSI Events." By doing so, we aimed to capture variability across subgroups that might be masked in a pure random sample.



Results Analysis: Summary Statistics, Skewness, and Kurtosis

After calculating summary statistics (mean, median, standard deviation, and variance) and visualizing them for both the random and stratified samples, we observe several key differences:

1. **Patient Occurrences**:
    - **Mean**: The random sample's mean was 25.41, compared to the stratified sample's mean of 17.68. This discrepancy indicates that the random sample may have captured more outliers with high values, raising the mean.
    - **Median**: The median was also higher in the random sample (11) than in the stratified sample (9), further reflecting that the random sample may have had more extreme values.
    - **Standard Deviation and Variance**: The random sample had a standard deviation of 41.83 and a variance of 1749.83, whereas the stratified sample's standard

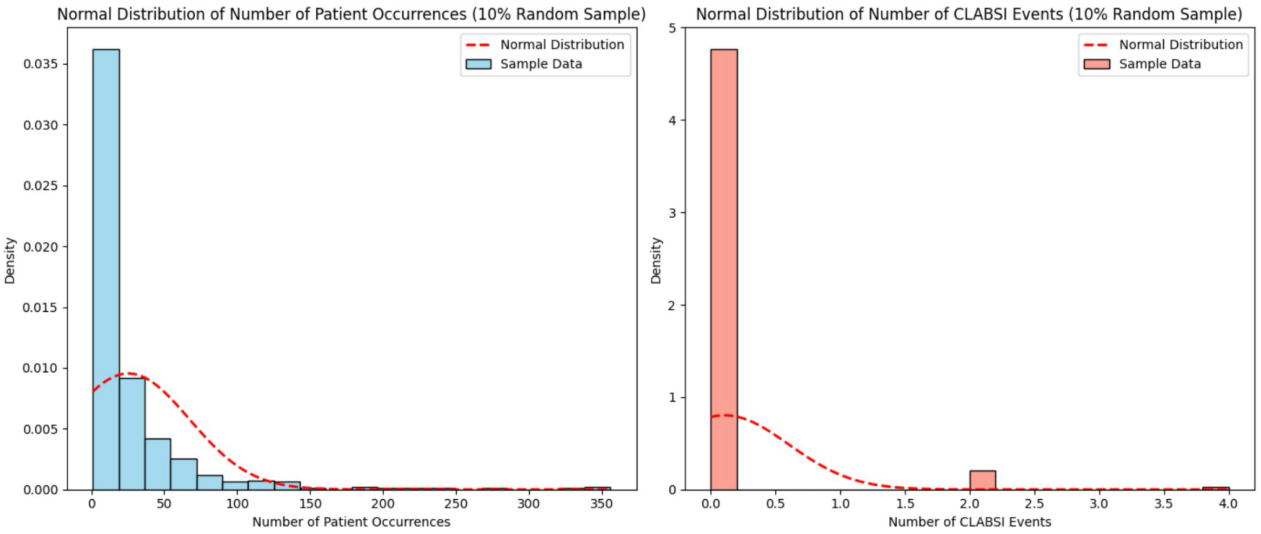
deviation and variance were much lower at 21.45 and 459.93, respectively. This lower variability in the stratified sample suggests it is more stable and representative of each subgroup.

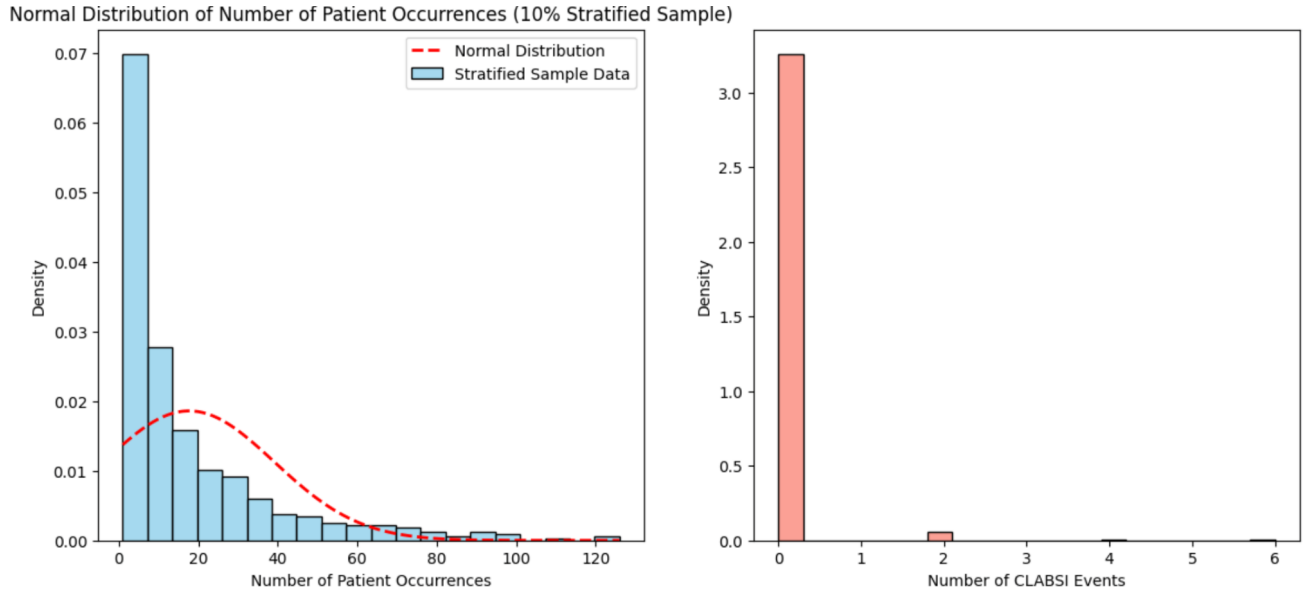
- **Skewness**: Skewness measures the asymmetry of a distribution. A positively skewed distribution indicates a longer right tail, often due to high outliers. The random sample showed greater positive skewness, likely caused by high outlier values in "Number of Patient Occurrences." The stratified sample had a lower skewness value, suggesting fewer outliers and a more balanced representation of the population.
- **Kurtosis**: Kurtosis measures the "tailedness" of a distribution, with high kurtosis indicating heavy tails (more extreme outliers) and low kurtosis indicating light tails. The random sample exhibited higher kurtosis, which, along with the skewness, suggests that it captured more extreme values. In contrast, the stratified sample showed a lower kurtosis, indicating fewer extreme values.

2. **CLABSI Events**:
    - **Mean**: The mean for "CLABSI Events" was 0.106 in the random sample, compared to a lower mean of 0.055 in the stratified sample, indicating that the random sample may have picked up more rare, high event counts.
    - **Median**: Both samples had a median of 0, meaning the majority of records had no CLABSI events. This suggests a high concentration of zero values in both samples, common in data with rare events.
    - **Standard Deviation and Variance**: The standard deviation was higher in the random sample (0.497) than in the stratified sample (0.414), and similarly, the variance was higher in the random sample (0.25 vs. 0.17 in the stratified sample). This increased variability in the random sample indicates a wider range of values, which is less controlled than the stratified approach.
    - **Skewness**: The positive skewness in both samples reflects the right-skewed nature of the data, with most occurrences around zero and few higher values. However, the skewness in the stratified sample was lower, suggesting a more balanced distribution across categories.
    - **Kurtosis**: Higher kurtosis in the random sample again indicates the presence of more extreme values or outliers in "CLABSI Events," while the stratified sample's lower kurtosis points to a less extreme distribution.

Visualization Insights

Bar plots comparing the mean, median, standard deviation, and variance across the random and stratified samples show that the random sample generally captures a broader range of values, resulting in higher variability (as seen in the standard deviation and variance). The stratified sample, on the other hand, provides lower variability across both variables, suggesting a more stable and controlled sampling approach. The higher kurtosis and skewness in the random sample visualizations reinforce that it includes more extreme values, or outliers, compared to the stratified sample.





Comparison of Results:

1. Representation of Rare Events:

- **Stratified Sampling**: Deliberately ensures that all relevant subgroups, including rare but significant CLABSI occurrences, are represented. By stratifying based on the 'CLABSI EVENT' variable, this method allows rare events to be adequately included in the sample. As a result, it provides more balanced data that represents the full range of CLABSI occurrences within the population.
- **Simple Random Sampling**: Random sampling does not specifically aim to capture rare events, as selection is purely random and does not guarantee coverage of each subgroup. This can lead to underrepresentation of CLABSI events, particularly if these occurrences are sparse or distributed unevenly across the population.

2. Distribution Characteristics:

- **Stratified Sampling**: Achieves a balanced distribution by ensuring representation from each subgroup based on CLABSI event occurrence. This balanced approach reduces variability within each subgroup, leading to a more consistent and representative dataset for analysis. The stratified sample reflects a diverse range of patient outcomes, which is beneficial when modeling complex healthcare scenarios.
- **Simple Random Sampling**: Due to the random selection, the sample often mirrors the natural skewness of the original dataset, where the majority of patients may have few or no CLABSI events, and only a small fraction experience frequent occurrences. This can result in a sample that lacks representation of the critical outlier cases, potentially skewing the model's predictions toward more common outcomes.

3. Predictive Accuracy and Model Performance:

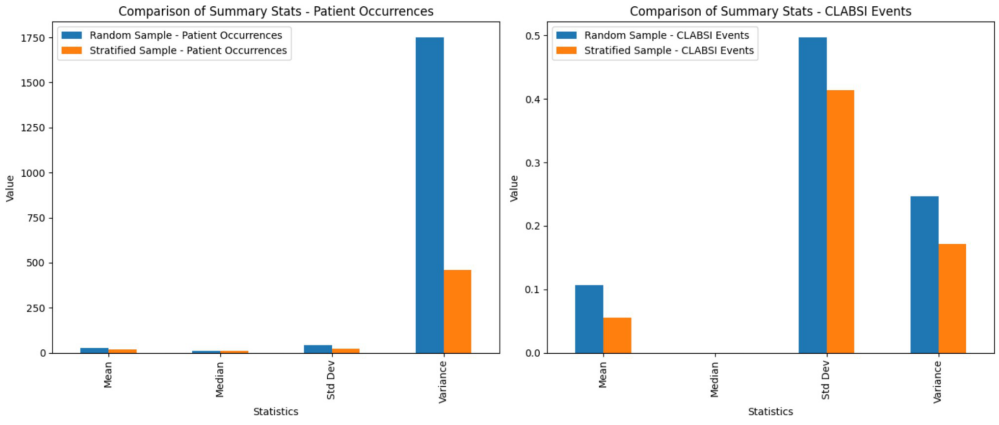
- **Stratified Sampling**: The balanced representation in stratified sampling improves the predictive accuracy of models by including data from both low- and high-risk groups for CLABSI events. This method produces a more reliable model, especially for predicting rare, high-stakes cases in a clinical context, such as patients at higher risk of CLABSI. Thus, stratified sampling supports the development of a model with strong performance across the entire spectrum of patient profiles.
- **Simple Random Sampling**: The model trained on a simple random sample may generalize well for common cases but could perform poorly for rare, high-risk cases due to the lack of sufficient data on these instances. Consequently, the model’s predictions may be less reliable in scenarios where rare events play a critical role, limiting its utility in high-stakes medical decision-making.

4. Generalizability of the Model:

- **Stratified Sampling**: Models based on stratified samples are generally more robust and generalizable, as they are trained on data that represents the entire population, including rare and extreme cases. The diversity of the sample improves the model's ability to generalize to new data, which is critical in healthcare settings where patient characteristics vary widely.
- **Simple Random Sampling**: Although simple random sampling is often generalizable, its lack of rare event representation means the model may struggle to accurately predict outcomes in high-risk or uncommon situations. The resulting model might be biased toward the majority group, thereby lacking sensitivity in detecting and responding to outlier cases.

5. Efficiency and Ease of Implementation:

- **Stratified Sampling**: While more complex to set up and requiring prior knowledge of the population's structure, stratified sampling is highly efficient when applied to smaller datasets, as it reduces variance and provides a balanced sample. However, it requires additional resources and planning to properly define and implement strata.
- **Simple Random Sampling**: Simple random sampling is straightforward and requires less effort to implement, as it does not necessitate prior knowledge of the population structure. It is a faster and easier method, particularly for large datasets. However, this efficiency comes at the potential cost of representational accuracy, especially for datasets with significant subgroup diversity.



<u>Conclusion</u>:

Our analysis of simple random and stratified sampling reveals distinct benefits for understanding CLABSI risk. Simple random sampling gave an unbiased overview of the population, including diverse cases and high-value outliers, but the added variability could impact model reliability. In contrast, stratified sampling provided a balanced representation of key patient groups, reducing the impact of outliers and offering a more stable foundation for focused analysis.

Together, these methods offer a complementary view: simple random sampling highlights overall

trends, while stratified sampling provides precision within patient subgroups. Using both approaches lays a strong foundation for building a predictive model that supports more accurate CLABSI prevention and ultimately enhances patient care.