Central Line-Associated Bloodstream infections (CLABSI) in California Hospitals Report

Kessie SHEN

Introduction

The National Healthcare Safety Network (NHSN), under the U.S. Centers for Disease Control and Prevention (CDC), defines central line-associated bloodstream infections (CLABSI) as laboratory-confirmed bloodstream infections occurring within 48 hours after central venous catheter placement, with no other identifiable sources of infection. These infections typically occur in patients undergoing treatment with a central line.

The mechanisms of CLABSI primarily include:

1.Infection at the catheter insertion site: Microorganisms, such as Staphylococcus aureus or Staphylococcus epidermidis residing on the skin, may enter through the puncture site or the surface of the catheter, becoming the initial source of infection. 2.Inadequate sterile techniques during catheter insertion: This can lead to direct microbial entry into the bloodstream. 3.Contamination of the catheter lumen: During catheter use, contamination may occur at infusion line connection points, medication injection sites, or other access points. In addition, patient-related factors can contribute to CLABSI. However, this analysis primarily focuses on external medical procedural factors, such as prolonged use of central venous catheters, improper catheter maintenance (e.g., contaminated dressings or delayed catheter replacement), and repeated catheter insertions. These factors can often be inferred based on hospital type, size, and the level of medical advancement in different regions.

According to the NHSN's definition, CLABSI is determined based on laboratory-confirmed bloodstream infections rather than clinical presentation. This may lead to cases of bloodstream infections caused by other sources being misclassified as CLABSI. This definition prioritizes monitoring purposes rather than accurately reflecting the actual clinical etiology, potentially overestimating the true incidence of CLABSI. Additionally, differences in how healthcare facilities implement the NHSN definition may affect data comparability.

The datasets I selected cover various types of hospitals, including acute care hospitals, critical access hospitals, long-term acute care hospitals, free-standing rehabilitation hospitals, and acute rehabilitation units reporting separately. They allow for an in-depth analysis of CLABSI rates and influencing factors across different healthcare settings and regions. This information is crucial for understanding infection trends, identifying risk factors, and improving infection prevention strategies

Methods

How and where the data were acquired.

This public access dataset was created by the California Department of Public Health (CDPH),can be found on the California Health and Human Services (CHHS) Open Data Portal.It offers comprehensive information on central line-associated bloodstream infections (CLABSI) reported by California hospitals via the CDC's National Healthcare Safety Network (NHSN). The data spans various hospital categories, including acute care hospitals, critical access hospitals, long-term acute care hospitals, free-standing rehabilitation hospitals, and acute rehabilitation units. Key variables include the number of infections, central line-days, standardized infection ratios (SIRs), and hospital-specific details. These features enable in-depth analyses of CLABSI trends, regional disparities, and variations across healthcare settings, making this dataset a crucial tool for advancing infection prevention and control efforts.

Cleaned and wrangled the data

Understand the Data Structure and Content

· Review Data Sources Since i m working with data spanning from 2013 to 2023,Before merging datasets from different years, I need to ensure that the information recorded in each dataset is consistent and aligned with the question I'm interested in. I also need to identify which variables are relevant to my analysis and which ones can be temporarily removed without affecting the results.

· Check Time Consistency: Make sure the data is consistent in terms of time intervals.When checking time consistency, only the year 2020 was split into two datasets, one for the first half and one for the second half. As a result, the month is shown as 6, while most other years have data for the full 12 months.

· Standardize Column Names Columns should be consistently named across datasets to facilitate merging and analysis.The column name for the 2013 table is "Observed_infections," although it represents the same data as "Number of reported CLABSIs," which is referred to

as “Infections_Reported” in other tables. A similar issue occurs with “Infections_Predicted,” and for one year, the column name is “Observed_Infections.” Additionally, there is a difference in case for “Central_line_days,” which appears as “Central_Line_Days” and “Central_line_Days” in 2021.

· Using the function:matches(“(?i)”)makes these matches more robust by ignoring case differences.

· Removing Unnecessary Columns:I understand that removing this field may reduce flexibility. However, due to inconsistencies in naming conventions—such as Facility_Name being used from 2017 to 2023 and Facility_Name1/2 from 2013 to 2016—merging these fields would result in numerous NA values, which could impact the overall clarity of the dataset. After carefully evaluating its relevance and confirming that its removal has minimal impact on the analysis of infection rates, I have decided not to retain this field in the current study.

Tools used for data exploration.

· ggplot2: This package was used for visualizing the CLABSI rates by county with boxplots, as well as for creating bar charts to display the count of facilities by type. · corrplot: This package was used to visualize the correlation matrix, providing insights into the relationships between numeric variables in the dataset. · lm(): The base R function lm() was used to perform linear regression, exploring how hospital size (as represented by bed count) influences the CLABSI rate.

Data Dictionary.after reading the Description Data Dictionary I believe that “Facility ID” is primarily a unique identifier used to track information about hospitals or healthcare facilities in the California Department of Public Health (CDPH) Electronic Licensing Management System (ELMS).

It is important for data management and record keeping, but it has little impact on the analysis of central line-associated bloodstream infection (CLABSI) data itself and does not directly affect the analysis of infection rates or hospital performance.

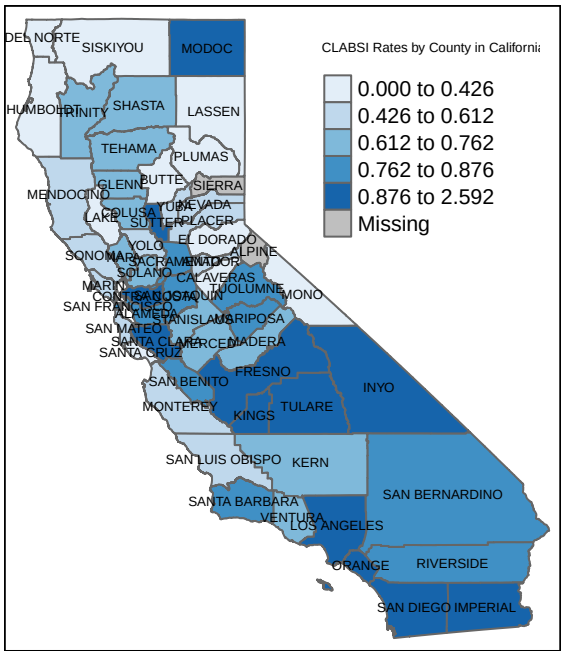
Therefore, I deleted it during the merging process. Clip_Adherence_Percent was reported in only four years, and On_Track_Toward_2020_Goal appeared only in 2017 and 2018, resulting in insufficient data. For this analysis, I have temporarily removed them.

Table 1: Preview of Cleaned Data

Year	State	County	InfectNumber	Pred_Infect	CentralLINE_Days	SIR	SIR_CI_95_Lower_Limit	SIR_CI_95_Upper_Limit	Comparison	Notes	Facility_Category	Facility_Type	SIR_2015	Months	Met_2020_Goal
2013	California	Santa Cruz	4	10.35	7353	0.3900000	0.1200000	0.930000	Low		NA	NA	1.093232	10.92853	NA
2013	California	Kings-Fresno	4	7.48	5791	0.5400000	0.1700000	1.290000	No difference	† See data dictionary	NA	NA	1.093232	10.92853	NA
2013	California	Fresno	0	0.00	4	0.8469149	0.3174618	3.328801	Too few to calculate		NA	NA	1.093232	10.92853	NA
2013	California	Orange	13	14.39	8191	0.9000000	0.5000000	1.510000	No difference		NA	NA	1.093232	10.92853	NA
2013	California	Alameda	6	20.33	11238	0.3000000	0.1200000	0.610000	Low	† See data dictionary	NA	NA	1.093232	10.92853	NA
2013	California	Alameda	0	1.38	874	0.0000000	0.0000000	2.160000	No difference		NA	NA	1.093232	10.92853	NA

Results

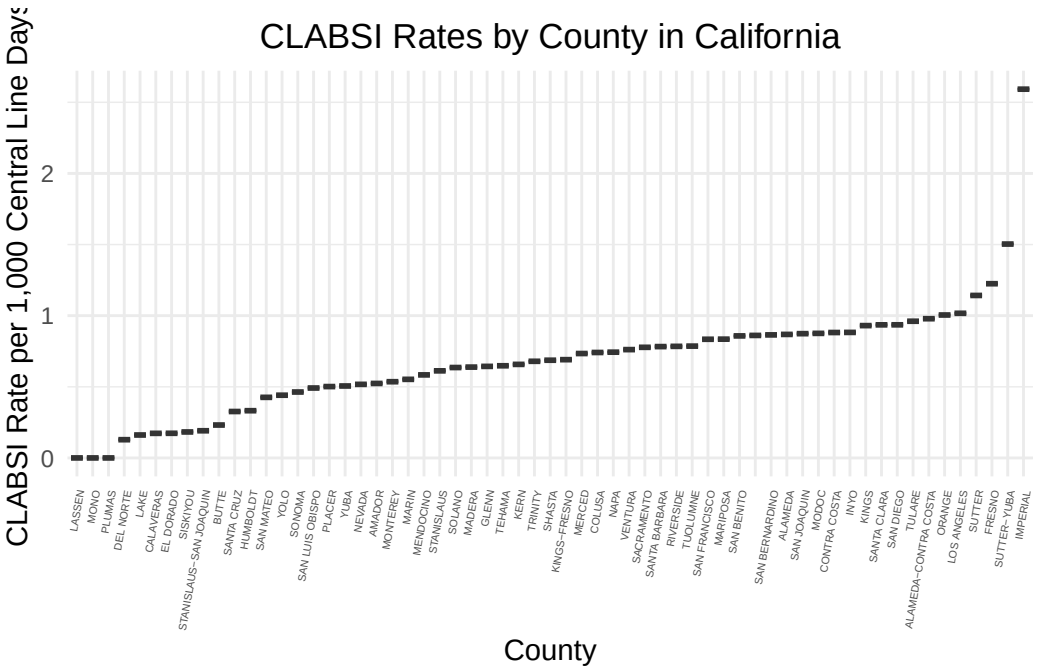
CLABSI infection rates between different regions in California.



The map uses color shading to represent the infection rates, with lighter colors indicating lower infection rates and darker shades indicating higher infection rates. The data is normalized by the number of central line days, and counties with missing data are highlighted in gray. This map allows for a quick geographic comparison of infection rates across the state, making it clear which regions have higher infection rates and which areas may need more targeted interventions.

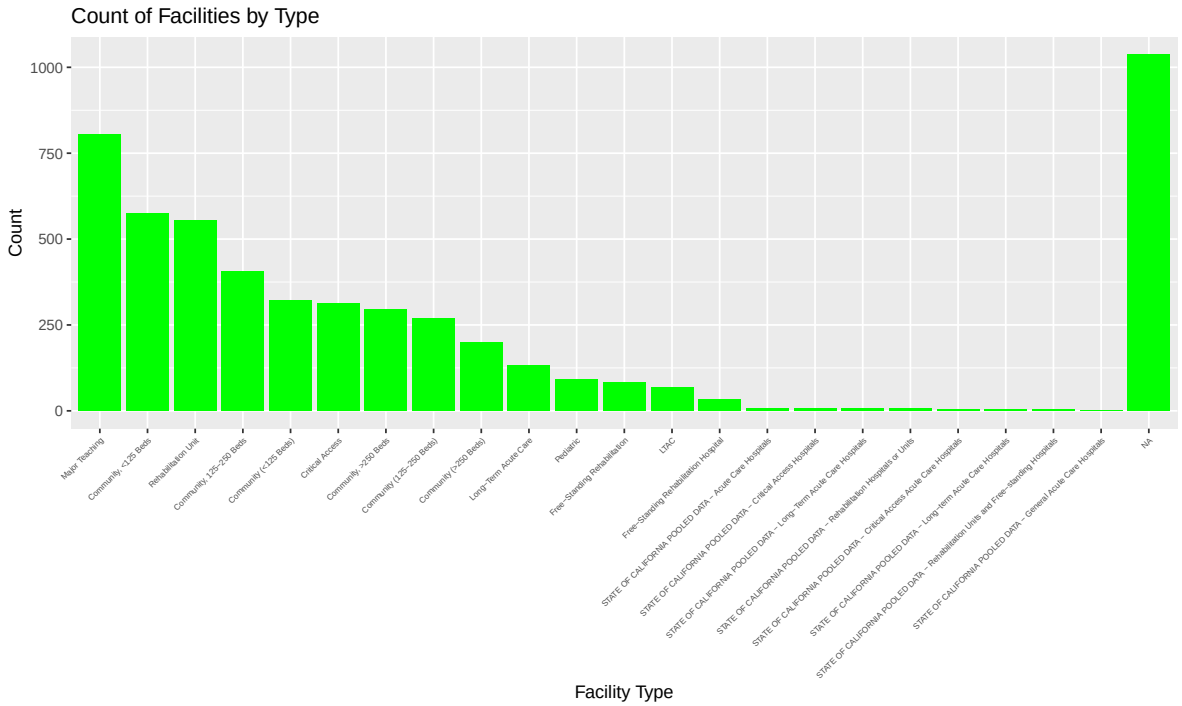
Urban vs. Rural: · Urban counties such as Los Angeles, San Francisco, and San Diego tend to have higher patient volumes, which could contribute to higher infection rates due to more frequent use of central lines and higher patient turnover.

· Rural counties, such as those in the Central Valley or northern regions like Del Norte and Siskiyou, might have lower infection rates due to fewer central line procedures or better infection control practices.



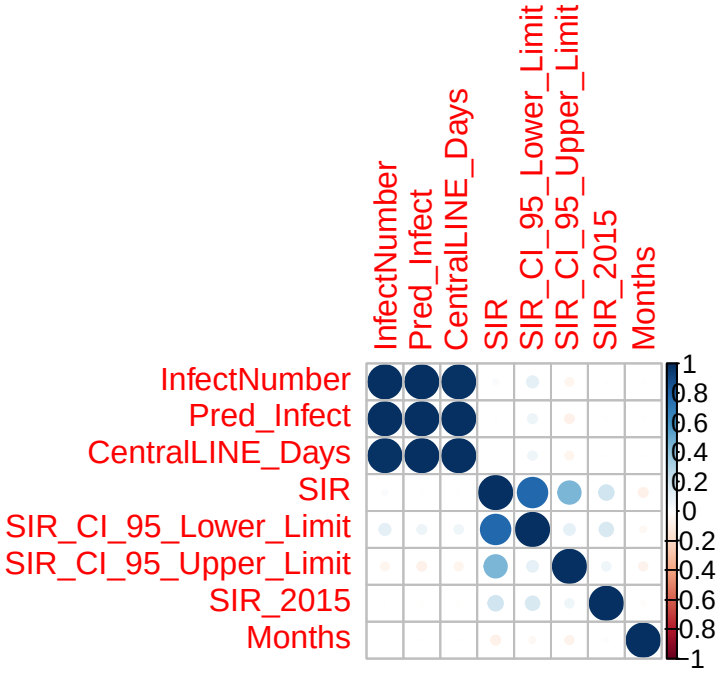
The box represents the interquartile range (IQR), where 50% of the counties' CLABSI rates fall. Outliers: Any counties with unusually high or low infection rates are marked as outliers and are located outside the whiskers. These are the counties with extreme CLABSI rates that deviate from the general trend.

Count of Facilities by Type



This bar chart illustrates the count of healthcare facilities in California by facility type.This chart shows that "Major Teaching" hospitals are the most common in California, followed by "Community" hospitals of varying bed sizes. Some specialized facilities, like LTACs and Pediatric hospitals, have much fewer facilities in comparison. It also highlights some data issues, as categories with names like "STATE OF CALIFORNIA POOLED DATA" and "NA" show up at the far right, which could be data inconsistencies or errors that need to be addressed.

The relationships between various numerical variables in the cleaned dataset.



This visualization is a correlation matrix that shows the relationships between various numerical variables.It helps in understanding the relationships between different variables related to hospital infection rates, central line days, and hospital characteristics.

Strong Positive Correlations Between InfectNumber, Pred_Infect, and CentralLINE_Days:

· Infections and central line days are directly related. Hospitals with more central line days are more likely to report more infections because the longer a central line is in place, the higher the risk of a bloodstream infection

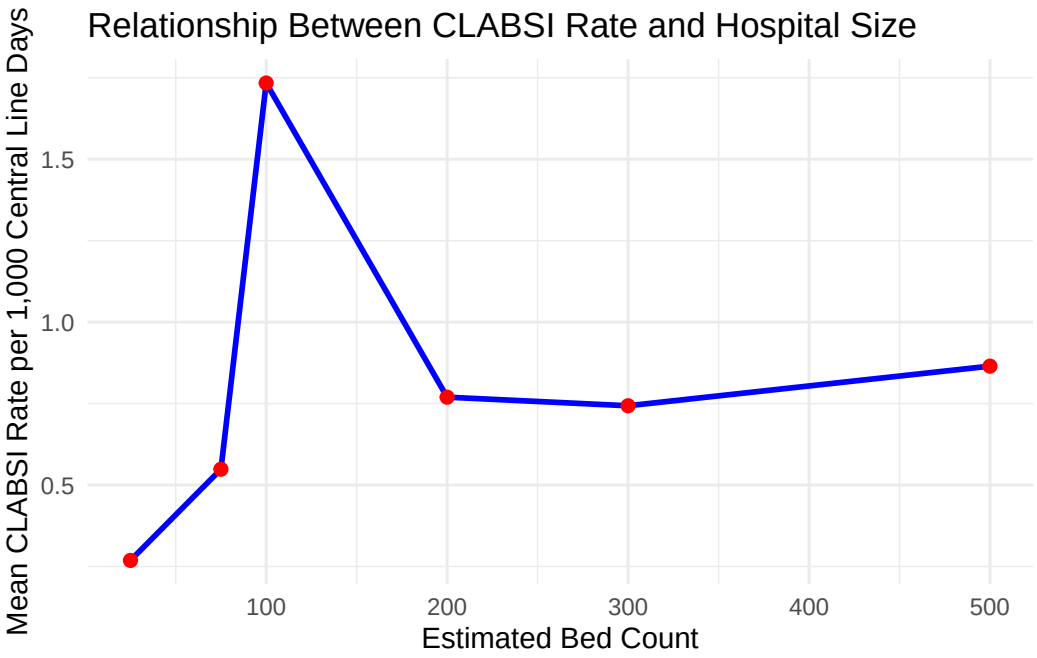
· As the number of reported infections (InfectNumber) increases, the Pred_Infect (the predicted infections based on baseline data or models) also tends to increase. This makes sense because the model predicting infections is likely using actual infection data as part of its calculation.

Moderate Correlations Between SIR and CLABSI_Rate: · SIR (Standardized Infection Ratio) is a metric that compares the number of infections in a healthcare facility to the number of infections expected based on national or baseline data. If the SIR is 1, the infection rate is as expected. If it's greater than 1, the facility has more infections than expected, and if it's less than 1, the facility has fewer infections than expected.

· CLABSI_Rate refers to the rate of central line-associated bloodstream infections in a given healthcare setting, usually expressed as the number of infections per 1,000 central line days.

· Increased infection rates are often associated with higher SIR values. If a hospital experiences a higher than expected rate of CLABSI, the SIR will be higher (above 1), which indicates worse performance in infection control.

The relationship between the incidence of CLABSI and the size of the hospital



This line chart illustrates the relationship between hospital size, represented by estimated bed count, and the mean CLABSI rate per 1,000 central line days. The chart shows that smaller hospitals (<100 beds) tend to have a gradual increase in CLABSI rates, peaking at facilities with around 100 beds. Medium-sized hospitals (200 beds) demonstrate a significant decrease in infection rates, potentially reflecting better resource allocation and infection control practices. Larger hospitals (300–500 beds) exhibit relatively stable infection rates, with a slight upward trend at the highest capacity levels, possibly due to the complexity of cases handled in such facilities. This trend highlights the varying challenges and strengths in infection prevention across different hospital sizes.

Potential Explanations 1.Smaller hospitals may have higher infection rates due to limited resources or suboptimal sterile practices.

2.Medium-sized hospitals may benefit from better management and resource allocation, leading to lower infection rates.

3.Larger hospitals, despite having advanced facilities, may face higher infection risks due to the complexity of cases.

Conclusion and Summary.

Understanding these correlations can help target specific variables for intervention, such as reducing unnecessary central line use.Infection prevention strategies may need to be tailored to the specific challenges faced by smaller and larger hospitals.

· Smaller hospitals may face resource limitations, while larger hospitals encounter challenges related to case complexity, both affecting CLABSI rates.

· Longer central line use directly correlates with higher infection risks.Predicted infections (Pred_Infect) and actual infections (InfectNumber) align closely, indicating a reliable infection prediction model.

· Categories like "NA" and "STATE OF CALIFORNIA POOLED DATA" highlight the need for standardized data cleaning processes to ensure accurate analyses.

Implications for Future Research and Practice 1.Targeted Infection Control: Interventions should address the specific needs of hospitals based on their size, type, and geographic location. 2.Policy Recommendations: Allocate resources and training to regions and facilities with higher CLABSI rates to improve outcomes

3.Data Improvements: Enhance data collection processes to ensure consistency and eliminate ambiguities in facility categorization.