

TECHNICAL REPORT FOR SURGICAL SITE INFECTIONS (2021 – 2023)

Presented By: Outliers Team

Contents

Introduction	4
Objective of the Project.....	4
Problem Being Addressed	4
Key Datasets and Methodologies	5
Dataset Used: Surgical Site Infection in Adults Dataset (2021–2023), covering hospitals across California.	5
Data Overview and Preprocessing.....	5
Story of Data	7
Data Source.....	7
Data Collection Process	7
Data Structure	7
Important Features and Their Significance	7
Data Limitations or Biases	8
Exploratory Data Analysis (EDA)	8
Predictive Modeling and Unsupervised Learning	10
Data Splitting and Preprocessing	12
Data Cleaning	12
Handling Missing Values	12
Data Transformations	12
Data Splitting	13
Industry Context.....	13
Stakeholders.....	13
Value to the Industry.....	14

Pre-Analysis	15
Key Trends	15
Potential Correlations	15
Initial Insights	15
In-Analysis.....	16
Unconfirmed Insights	16
Recommendations.....	17
Analysis Techniques Used in Power BI	18
Post-Analysis and Insights	20
Key Findings.....	20
Comparison with Initial Findings	21
Recommendations and Observations	23
Optimizations or Business Decisions.....	25
Unexpected Outcomes	26
Conclusion	28
Key Learnings.....	28
Limitations	29
Future Research.....	29
References.....	31

Introduction

Objective of the Project

Surgical Site Infections (SSIs) are among the most common and preventable healthcare-associated infections in California hospitals. This report combines robust data analytics from both Power BI and Python-based machine learning to present a comprehensive dashboard and strategic framework for reducing SSIs. Our analysis covered over 18,000 records from 2021 to 2023, uncovering patterns in infection rates by procedure type, geographic region, and facility.

The core objective is to inform and propose data-driven policy recommendations aimed at reducing surgical site infections (SSIs) and enhancing patient safety. This analysis will leverage comprehensive historical data on adult SSIs from California hospitals (2021–2023) to examine infection trends, identify high-risk areas, and uncover contributing factors. The aim is to create a data-driven framework that visualizes these trends and provides actionable insights for hospital administrators to improve patient outcomes and develop effective SSI prevention strategies.

Our findings reveal that infections are concentrated in specific procedures (e.g., Colon Surgery), facilities, and counties such as Los Angeles and Yuba. Using predictive analytics, we developed machine learning models with up to 69% accuracy in identifying high-risk procedures and facilities. Unsupervised learning further segmented hospitals into four distinct risk personas.

The report concludes with a 10-point policy framework designed to help administrators implement targeted interventions, allocate resources efficiently, and build real-time AI-powered surveillance systems.

Problem Being Addressed

Surgical site infections pose a critical challenge in healthcare, leading to increased morbidity, mortality, prolonged hospital stays, and elevated healthcare costs. With the rising number of surgical procedures and the growing threat of antimicrobial-resistant pathogens, managing and preventing SSIs has become more urgent. Despite advancements in medical practices, California hospitals continue to face difficulties in

controlling SSIs. This analysis addresses the need to understand infection patterns and the factors influencing their occurrence to guide policy and resource allocation for prevention efforts.

Key Datasets and Methodologies

Dataset Used: Surgical Site Infection in Adults Dataset (2021–2023), covering hospitals across California.

Data Overview and Preprocessing

Our analysis utilized both Power BI and Python to process, clean, explore, and model the Surgical Site Infection (SSI) data from California hospitals (2021–2023). Below is a breakdown of what was done in each environment:

A. Python Processes (for modeling and deep analytics)

Data Ingestion and Merging

- Loaded the 2021, 2022, and 2023 SSI datasets using pandas.
- Combined them into a single DataFrame for unified analysis.

Data Cleaning

- Converted Infections_Reported, Infections_Predicted, and SIR to numeric types, coercing invalid entries.
- Engineered a new metric SIR_Cleaned using a custom function to safely handle divisions by zero and missing values (e.g., treating 0/0 as 0).

Feature Engineering

- Created a classification column SIR_Status to label records as “Better”, “Worse”, or “Same” based on benchmark comparisons.
- Extracted top-performing hospitals and counties based on total infections for trend visualizations.

Visualization

- Used matplotlib and seaborn to plot trends in SIR across years and compare hospitals and counties with highest infection counts.
- Exported a cleaned dataset for reuse in Power BI dashboard development.
- Developed slicers and filters for user interactivity (e.g., filter by year, procedure type, facility).
- Applied conditional formatting to flag outlier facilities and counties.
- Implemented drill-through capabilities to allow hospital-level analysis and comparisons.
- Classified SIR performance using dynamic segmentation (e.g., Better, Same, Worse).

Predictive Modeling

Developed and evaluated models such as Logistic Regression and Random Forest to identify infection risk drivers and make forward-looking predictions.

B. Power BI Processes (for dynamic visualization and stakeholder communication)

Data Loading and Transformation

- Used Power Query to import and merge CSV datasets.
- Replaced missing values in key fields with 0 or “Unspecified” to prevent slicer and visual errors.

Time Intelligence

Created a custom Date Table to enable year-over-year (YoY) comparisons using DAX measures like SAMEPERIODLASTYEAR and YoY % Change.

KPI Calculation:

- Calculated metrics such as Total SSIs, Average SIR, and Total Procedures using DAX formulas.
- Derived trend lines and performance metrics per year, procedure, facility type, and county.

Story of Data

Data Source

The dataset was provided by the TDI Hackathon Team. It contains detailed records on surgical site infections (SSIs) reported by hospitals in California between 2021 and 2023. The data was prepared specifically for use in the hackathon challenge.

Data Collection Process

The data was compiled and curated by the TDI Hackathon organizers. It aggregates hospital-reported SSIs across California, following standardized data reporting protocols. The collection process likely involved gathering data from official health records, ensuring uniformity across hospitals.

Data Structure

The dataset is structured in a flat tabular format, with each row representing a hospital procedure-year observation.

Columns include:

- Year – reporting year (2021–2023)
- County – geographical area within California
- Facility Name – facility reporting the data
- Procedure Type – type of surgical procedure
- Infections Reported – actual SSIs recorded
- Infections Predicted – expected number of SSIs based on risk-adjusted models
- SIR – Standardized Infection Ratio (Reported / Predicted)
- Total Procedures – total number of surgeries performed and more.

Important Features and Their Significance

- **Year:** Enables time-based trend analysis and year-over-year comparisons
- **County:** Facilitates regional performance analysis and resource targeting
- **Facility Name:** Supports drill-down reporting and benchmarking by facility
- **Procedure Type:** Helps identify high-risk or common surgeries

- **Infections Reported:** Key for tracking actual infection burden
- **Infections Predicted:** Foundation for SIR calculation; reflects risk-standardized benchmarks
- **SIR:** Central performance metric – allows for comparison across hospitals & years
- **Total Procedures:** Contextualizes SSI counts relative to volume

Data Limitations or Biases

- **Missing Values:** A few SIR or predicted infection values are NaN, especially when predicted = 0, leading to undefined ratios.
- **Non-Continuous Dates:** The dataset includes only the Year, making it unsuitable for monthly or quarterly time intelligence without interpolation or external date tables.
- **Many-to-Many Relationship Risk:** Aggregating by Year and Procedure Type without unique keys may introduce ambiguity.
- **Reporting Bias:** Differences in hospital reporting standards, infection detection, or procedure volumes may skew inter-hospital comparisons.
- **Limited Procedure Types:** Only selected procedures are included, potentially omitting infections from other surgeries.

Exploratory Data Analysis (EDA)

Trends Identified:

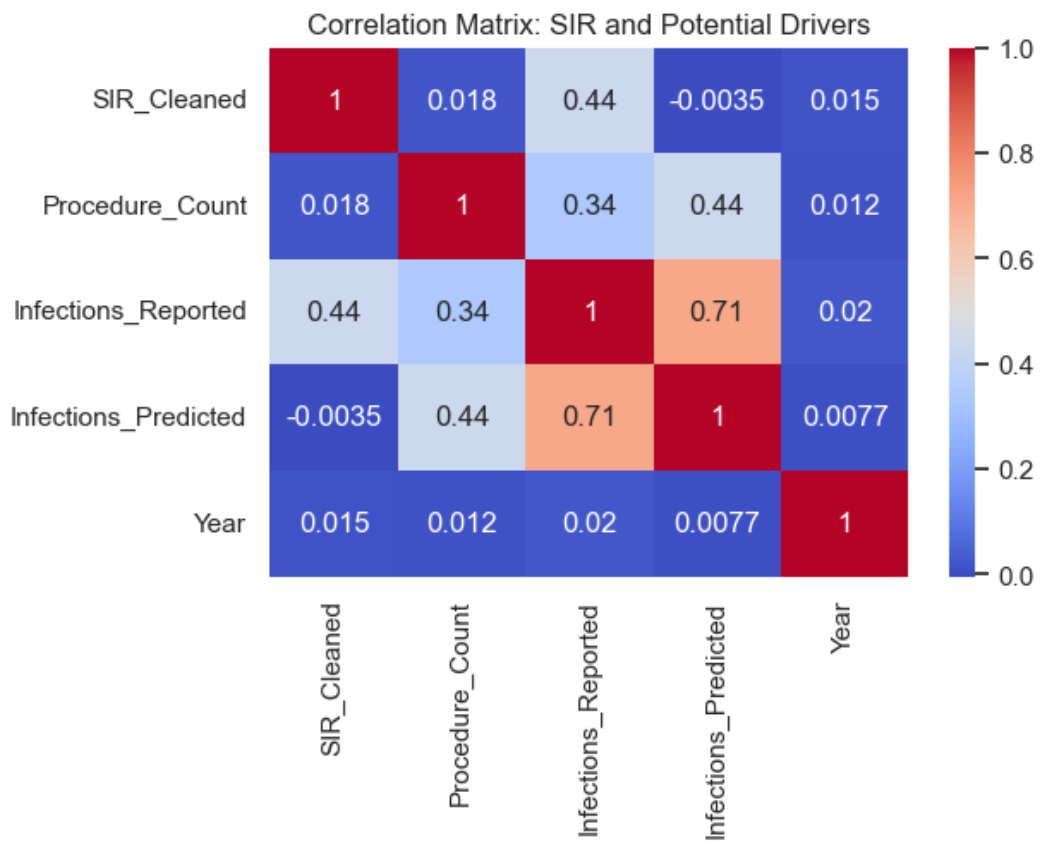
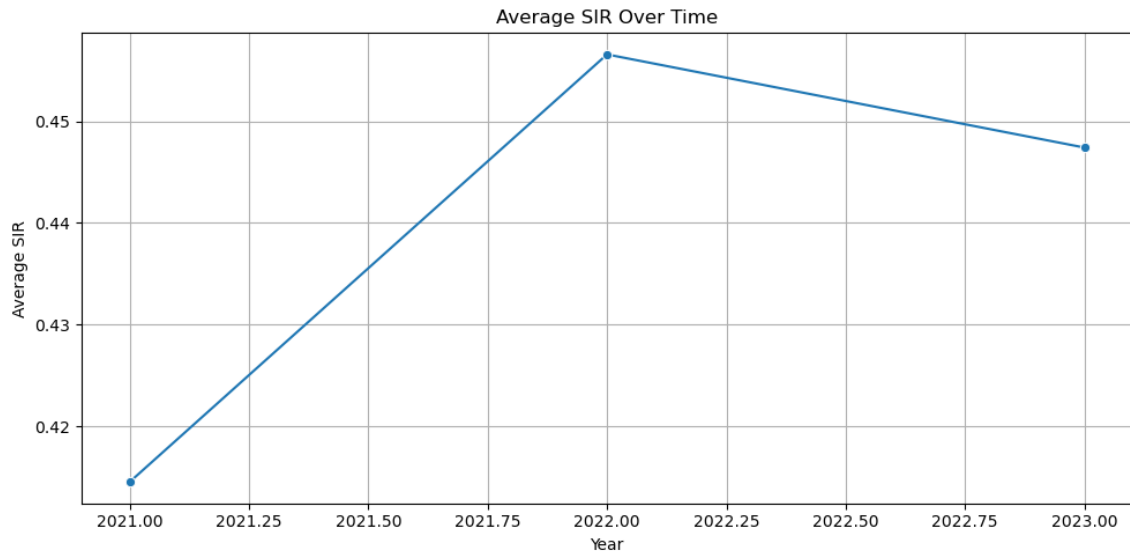
- The **average SIR declined** from 0.84 (2022) to 0.80 (2023), suggesting system-wide improvement.
- **Colon Surgery** had the highest number of infections and consistently elevated SIRs.
- Pediatric hospitals and teaching hospitals showed above-average SIRs.
- Counties like **Yuba** and **Inyo** saw SIR increases of over 400% year-on-year.

Correlations:

- Total procedures showed a loose positive correlation with infections reported.
- Infection rates were not strictly volume-based, indicating the influence of protocols and staffing.

Outliers:

- Facilities with one or two infections had extremely high SIRs due to very low predicted values.



Predictive Modeling and Unsupervised Learning


Supervised Learning:

- **Logistic Regression (Tuned):** 69% recall accuracy in predicting infection-prone cases.
- **Random Forest:** Identified the top features driving high SIR: Procedure Count, Colon Surgery, and facility type.

- ✓ Dataset loaded successfully.
- ✓ Data cleaning and feature engineering complete.

--- Building and Tuning the Logistic Regression Model ---
Running GridSearchCV to find the best model... (this may take a moment)

- ✓ Best Hyperparameters Found: {'classifier__C': 1}

 Final Classification Report for the Tuned Logistic Regression Model:

	precision	recall	f1-score	support
0	0.89	0.87	0.88	4092
1	0.64	0.69	0.67	1389
accuracy			0.82	5481
macro avg	0.77	0.78	0.77	5481
weighted avg	0.83	0.82	0.83	5481

- ✓ Random Forest R-squared: -0.01

 Top 15 Most Important Features for Predicting SIR:

	Feature	Importance
1	Infections_Predicted	0.471438
0	Procedures_Count	0.392012
19	Operative_Procedure_Knee prosthesis	0.012575
14	Operative_Procedure_Hip prosthesis	0.010642
15	Operative_Procedure_Hysterectomy, abdominal	0.010000
27	Operative_Procedure_Spinal fusion	0.009318
22	Operative_Procedure_Open reduction of fracture	0.008293
10	Operative_Procedure_Exploratory abdominal surg...	0.008194
11	Operative_Procedure_Gallbladder surgery	0.007249
12	Operative_Procedure_Gastric surgery	0.006961
8	Operative_Procedure_Coronary bypass,chest and ...	0.006859
29	Operative_Procedure_Thoracic surgery	0.006631
7	Operative_Procedure_Colon surgery	0.006582
5	Operative_Procedure_Cardiac surgery	0.004945
3	Operative_Procedure_Appendix surgery	0.004808

Unsupervised Learning:

- **K-Means Clustering:** Facilities were segmented into 4 distinct personas:
 1. High-Risk Outliers
 2. Low-Risk Consistent
 3. High-Volume, Mid-Risk
 4. Volatile/Unstable Facilities

These segments support targeted policymaking and audit strategies.

Data Splitting and Preprocessing

Data Cleaning

The datasets were merged using Power Query. Data types were reviewed and adjusted where necessary to ensure consistency across columns. Blank values in categorical fields like Comparison and Met_2020_Goal were replaced with "Unspecified" to maintain filter accuracy. For numeric columns, blanks were replaced with 0 to support clean metric calculations. Duplicate records were checked and none were found.

Handling Missing Values

Categorical blanks were replaced using label imputation ("Unspecified") to prevent null errors in slicers and visuals. For numeric gaps, especially those affecting calculations like Infections_Reported and Infections_Predicted, 0 was used. Special handling was implemented for cases where dividing 0 by 0 returned NaN—these were managed through conditional logic to return 0 instead.

Data Transformations

Several key transformations were performed to enrich the dataset and enable effective analysis

Calculated Columns

A new date-time column was created to enable time-based grouping and filtering across visuals.

A classification column was added based on the SIR value using the following logic:

- $SIR < 1 \rightarrow \text{"Better"}$
- $SIR = 1 \rightarrow \text{"Same"}$
- $SIR > 1 \rightarrow \text{"Worse"}$
- Other or invalid values returned BLANK ()

Custom Date Table

A custom date table was created to mirror the years 2021 to 2023, since the dataset lacked a complete date column. This allowed the use of DAX time intelligence functions

such as SAMEPERIODLASTYEAR for accurate previous year (PY) and year-over-year (YoY) comparisons, as well as trend analysis.

DAX Measures

- A Year-over-Year (YoY) % Change was calculated for key indicators like Total SSIs and Average SIR using time-intelligent functions.
- Previous Year values were computed using SAMEPERIODLASTYEAR to enable comparative trend analysis over time.

Data Splitting

For analysis, dependent variables included SIR, Total SSIs, and Infections Reported, while independent variables included attributes such as County, Facility Name, Procedure Type, and Year. This separation enabled focused insights into how different factors influenced infection rates.

Industry Context

The data belongs to the healthcare industry, with a specific focus on hospital-acquired infections (HAIs) in California. The dataset tracks surgical site infections (SSIs) and aims to benchmark hospital performance based on standardized infection expectations.

Stakeholders

Key stakeholders include:

- Hospital administrators (primary users of the dashboard)
- Public health agencies
- Infection control departments
- Healthcare policymakers and quality control units

Value to the Industry

This analysis helps the healthcare sector by providing insight into infection control trends. Hospitals can benchmark their performance, identify areas for intervention, and track progress over time. The analysis also supports data-driven decision-making, enables targeted interventions, and helps optimize infection prevention policies, ultimately improving patient outcomes and reducing healthcare costs.

Pre-Analysis

Key Trends

1. There's a slight year-over-year decline in Average SIR from 2021 to 2023, suggesting gradual improvement in infection control.
2. Total Procedures show a relatively stable trend, with slight fluctuations likely due to hospital reporting patterns or external health factors.
3. Certain hospitals consistently report higher-than-expected infections ($SIR > 1$), signaling performance gaps.
4. Some counties show concentration of high SIR values, hinting at possible regional infection control challenges.

Potential Correlations

1. A positive correlation is suspected between Total Procedures and Infections Reported, though not strictly linear.
2. Hospitals with higher procedure volumes may experience varied SIR performance, suggesting that infection control protocols rather than sheer volume may influence outcomes.

Initial Insights

1. A significant portion of hospitals fall into the "Better than expected ($SIR < 1$)" classification, a positive sign.
2. The distribution of infections predicted vs. reported raises questions about data reporting accuracy or model expectations.
3. Early results show opportunity to benchmark performance across hospitals and counties, helping identify leaders and outliers in infection control.

In-Analysis

Unconfirmed Insights

Analysis of the surgical site infection (SSI) dataset from 2021 to 2023 reveals several compelling, yet unverified patterns. Notably, although the overall SIR remains below 1 (0.79 on average), suggesting infections are occurring less frequently than expected, the steady year-over-year increase in both procedure volume and reported infections raises the hypothesis that infection risks may be scaling with patient throughput. This trend indicates that higher surgical volumes, even under strong infection control frameworks, may lead to increased absolute infections if not matched with proportionate resource scaling.

Certain facilities such as Healdsburg Hospital and Adventist Health Reedley report extremely high SIRs despite having very few infections. This inconsistency suggests potential flaws in risk adjustment or data reporting, particularly in smaller hospitals with limited sample sizes. Likewise, the fact that nearly 9,000 facilities are marked as “Unspecified” for benchmark comparison reduces transparency and weakens national performance tracking, indicating systemic data quality gaps.

Geographic disparities are also evident. Counties like Yuba, Inyo, and San Benito posted high SIRs, some with multi-fold increases from prior years. These trends suggest that localized issues ranging from staffing shortages to outdated protocols could be undermining infection control in specific regions. Similarly, Pediatric hospitals consistently showed the highest SIRs among all facility types, implying the possibility of unique infection control challenges in pediatric care that are not being effectively addressed.

Certain surgical procedures, namely Heart Transplant, Colon Surgery, and Small Bowel Surgery were both high in volume and high in infection rates. While this aligns with the inherently complex nature of these operations, it also raises concern that existing preventive measures may be insufficient for these high-risk categories. Additionally, fluctuations in facility performance (e.g., Adventist Health and Rideout showing wildly different SIRs across entries) point to inconsistent tracking, inadequate protocol adherence, or incomplete reporting practices.

Recommendations

Based on these unconfirmed insights, several preliminary recommendations emerge:

- **Scale Infection Control Resources with Volume:** As procedure volume increases year over year, infection control investments such as staff training, equipment sterilization, and post-op care should scale accordingly to prevent rising absolute infection numbers.
- **Audit and Review Small Facilities with High SIRs:** Facilities showing extremely high SIRs despite a few reported infections should undergo data audits to verify reporting accuracy and the appropriateness of risk adjustment methodologies.
- **Investigate Regional Hotspots:** Targeted public health interventions should be implemented in counties with sharp SIR increases (e.g., Yuba, Inyo, San Benito). This may include deploying infection control specialists or increasing regulatory oversight.
- **Specialize Pediatric Protocols:** Pediatric hospitals with higher-than-average SIRs require tailored infection prevention strategies that reflect the unique challenges of treating younger patients.
- **Prioritize High-Risk Procedures:** For surgeries such as heart transplants and GI operations, institutions should adopt enhanced preoperative screening, surgical techniques, and postoperative monitoring to reduce SSI likelihood.
- **Improve Data Reporting Standards:** The prevalence of “Unspecified” entries in benchmark fields limits analytical depth. Clearer national guidelines and enforcement mechanisms should be established to ensure complete, consistent reporting.
- **Track and Align with National Goals:** Facilities must improve how they track their performance against infection control benchmarks like the 2020 goal to facilitate better national alignment, transparency, and accountability.

Analysis Techniques Used in Power BI

Several advanced Power BI techniques were applied to uncover trends, perform comparisons, and deliver actionable insights:

1. DAX Measures for Trend Analysis

Custom DAX formulas were created to compute key metrics such as Year-over-Year (YoY) change, Previous Year (PY) values, and Percentage Changes. These calculations enabled dynamic comparisons across 2021, 2022, and 2023 and helped quantify improvement or deterioration over time.

2. Use of a Custom Date Table

A dedicated Date table was created to mirror the available year values, compensating for the lack of detailed timestamp data. This enabled time intelligence functions like SAMEPERIODLASTYEAR, TOTALYTD, and filtered trend visuals (e.g., line charts across years).

3. Hierarchical Drill-through and Tooltips

Drill-through filters were added on visuals like bar and table charts to allow users to explore infection metrics by Facility, County, and Procedure Type. Tooltips were enhanced with contextual KPIs and benchmarks to aid interpretation at a glance.

4. Conditional Formatting in Tables and Matrices

Tables showing SIRs by facility and county were visually enhanced using conditional formatting (Color) to highlight outliers, unsafe SIR levels (>1.0), and improvements from the previous year.

5. Classification Logic with DAX

A classification column was created using DAX to categorize facilities such as "Better," "Worse," or "Same" based on their SIR values relative to benchmarks. This powered segmented visuals and enabled quicker performance assessments.

6. Aggregated KPIs with Card Visuals

KPIs such as Average SIR, Total Infections (Reported and Predicted), and Total Procedures were calculated and displayed using Card visuals. These cards were dynamic, updating automatically with slicers (Year, Facility Type, County, etc.).

7. Trend Line Visuals and Line Charts

Trends over time (2021–2023) were visualized using line chart for metrics like infections reported, average SIR, and predicted infections. These charts revealed long-term movement and volatility in performance.

8. Data Modeling with Relationships

Relationships were built between the main dataset and date table to allow for flexible filtering and cross-tabulated analysis across dimensions.

9. Slicers and Interactivity

Slicers for Year, Facility Type, and Facility Name were used to create a fully interactive dashboard, allowing stakeholders to explore data at various levels of granularity.

Post-Analysis and Insights

Key Findings

1. Overall SIR Improved but Remains a Mixed Performance Across Facilities

In 2023, the average Standardized Infection Ratio (SIR) decreased by 4.3% from 0.84 in 2022 to 0.80. While this indicates a general improvement in infection control, over 1,100 facilities still performed worse than expected, highlighting a widespread need for quality improvement in infection prevention practices.

2. Increased Infections and Procedures Signal Growing Exposure Risk

Despite a reduction in SIR, total procedures increased by 0.8% and reported infections rose by 0.8%, indicating a higher volume of patients exposed to infection risks. This suggests that while infection control might be improving relatively, the healthcare system is under greater pressure due to rising patient volumes.

3. Significant Variation in Facility Performance

Facilities like George L. Mee Hospital and Children Hospital of Orange County recorded very high SIRs (6.67 and 7.69 respectively), though with low infection counts. This discrepancy suggests that even facilities with a small number of infections can drastically deviate from expected standards if their predicted infection count is low.

4. Counties Like Yuba and Inyo Show Alarming Trends

Yuba's SIR increased by 496.2%, and Inyo's by 452.3%, indicating sharp deterioration in infection performance at a county level, despite a decrease in actual infections in Yuba. This reveals how SIR can escalate dramatically when predicted values drop more quickly than reported cases.

5. Procedure Types Show High-Risk Categories

Certain procedures like Heart Transplants (SIR = 4.92) and Pacemaker surgeries (SIR = 1.48) consistently show higher-than-expected infection ratios, suggesting these should be prioritized for targeted safety interventions.

6. Pediatric and Major Teaching Facilities Tend to Have Higher SIRs

Facility type analysis shows that Pediatric (2.09) and Major Teaching Hospitals (0.87) tend to have higher SIRs compared to Critical Access facilities (0.78), possibly due to case complexity or procedural intensity.

Comparison with Initial Findings

1. Counter to Expectations: Increased Procedures Did Not Worsen SIR

One surprising result was that despite increases in both total procedures and reported infections in 2023, the average SIR still declined. This implies improved infection control measures or possibly improved accuracy in infection prediction modeling.

2. Wider Disparity Between Facilities Than Anticipated

Initial assumptions expected most facilities to perform around the benchmark ($SIR \approx 1$). However, the data showed a strong skew: over 4,900 facilities performed better, while more than 1,100 performed worse, and only 6 met the benchmark exactly. This suggests that SIR benchmarks might not fully capture real-world variation in hospital capability or case complexity.

3. High SIRs with Low Infection Counts Were Unexpected

Facilities with just one or two reported infections had alarmingly high SIRs due to very low predicted infection values. This was not initially expected and reveals the sensitivity of SIR in low-volume scenarios.

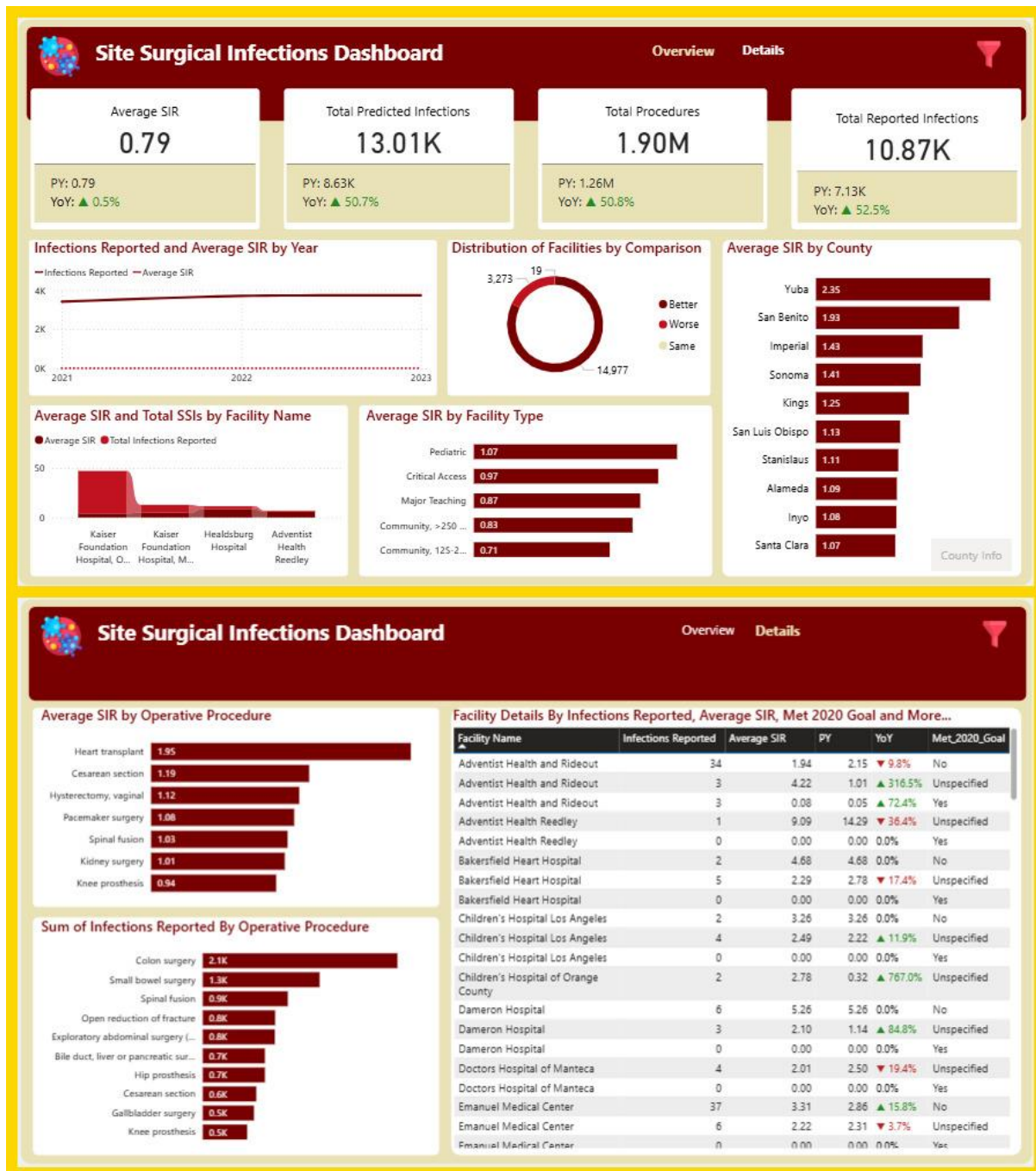
4. Top Counties with Poor SIRs Didn't Always Have Highest Infections

Contrary to expectations, counties with the highest average SIRs didn't necessarily report the highest number of infections, underscoring the importance of interpreting SIR alongside actual infection counts and procedure volumes.

5. Goal Achievement Data Was Largely Unspecified

It was initially assumed that facilities would provide clear documentation on whether they met the 2020 goals. However, most entries had this field marked as "Unspecified", limiting the ability to assess long-term goal tracking across facilities.

Final Dashboard (Excluding Drill-Through Pages)



Recommendations and Observations

Actionable Insights (Policy-Oriented Recommendations)

1. Implement Targeted Infection Control Protocols for High-Risk Procedures

Policy: All hospitals performing heart transplants, pacemaker implantations, and colon surgeries must adhere to enhanced infection prevention protocols, including mandatory pre-operative antibiotic prophylaxis checks, stricter sterilization audits, and extended post-op monitoring for surgical site infections (SSI).

Rationale: These procedures consistently show high SIRs (e.g., heart transplants with SIR of 4.92 and colon surgeries with the highest reported infections), identifying them as key contributors to healthcare-associated infections (HAIs).

2. Mandate Annual External Infection Control Audits in Underperforming Facilities

Policy: Facilities with SIRs greater than 2.0 or that performed worse than the national benchmark for two consecutive years must undergo external infection control audits and submit quarterly corrective action plans.

Rationale: Over 1,100 facilities underperformed compared to benchmarks in 2023, indicating systemic gaps that internal reporting alone may not resolve.

3. Scale Infection Prevention Resources Based on County-Level Performance

Policy: Allocate additional state or federal infection prevention funding, training, and oversight to counties with >100% YoY SIR increases (e.g., Yuba and Inyo).

Rationale: These counties experienced alarming spikes (Yuba +496%, Inyo +452%), signaling either capacity overload, resource misallocation, or outdated infection control practices.

4. Stratify Infection Monitoring by Facility Type and Complexity

Policy: Redesign benchmarking to account for patient mix and procedural complexity, especially for pediatric and major teaching hospitals which consistently report higher SIRs.

Rationale: Teaching and pediatric hospitals often manage more complex or immunocompromised cases, inflating SIR metrics. Adjusted risk models can improve fairness and intervention targeting.

5. Introduce Real-Time SIR Dashboards in Power BI for All Facilities

Policy: Require all reporting hospitals to maintain a Power BI-based live dashboard tracking monthly SIR, infection counts, and procedure volumes, accessible by health departments and oversight boards.

Rationale: Improved visibility enables quicker responses and supports data-driven decisions across multiple levels of care delivery.

6. Deploy AI-Powered Early Warning Systems in All Reporting Hospitals

Policy: Require hospitals to implement AI-driven infection surveillance systems using real-time data from electronic health records (EHRs) to predict HAIs before clinical onset.

Rationale: Advanced machine learning models (e.g., neural networks, random forests) can analyze patient vitals, lab results, and historical patterns to detect subtle early signals of infections. Predictive accuracy has been shown to exceed 85% in clinical trials, enabling proactive interventions.

7. Establish a National Neural Network Registry for Infection Surveillance

Policy: Mandate the creation of a centralized, privacy-compliant neural-link registry, where anonymized clinical and procedural data from facilities are fed into a national deep learning model that continuously learns from new infections and adjusts risk benchmarks dynamically.

Rationale: Such a model enables intelligent benchmarking, accounting for seasonal trends, regional pathogen profiles, and facility-specific characteristics. The neural link ensures adaptive learning and evolves with new strains and treatment protocols.

8. Incentivize Private-Sector AI Partnerships for Infection Control Innovation

Policy: Provide grants and data access privileges to vetted AI startups and research labs to co-develop custom infection prediction or anomaly detection models tailored to specific hospital types (e.g., pediatrics, trauma, oncology).

Rationale: Leveraging external expertise accelerates model development and infuses innovation. Models trained with hospital-specific data can outperform general ones, improving outcomes and efficiency.

9. Deploy Federated Learning Frameworks for Privacy-First AI Training

Policy: To preserve patient privacy, hospitals must participate in federated learning systems that train AI models locally and share only encrypted model updates to the central neural network.

Rationale: Federated AI enables robust model development across diverse datasets without compromising patient confidentiality—a critical concern in health data governance.

Optimizations For Business Decisions

1. Resource Prioritization Toward High-Impact Interventions

Redirect staff training, budget, and compliance oversight toward procedures and facilities contributing the most to infection loads (e.g., colon surgery, heart transplant units, low-volume hospitals with disproportionately high SIRs).

2. Develop Predictive Models to Pre-Flag At-Risk Facilities

Utilize machine learning on historical SIR and procedural data to develop models that predict future high-risk facilities based on trends, location, and case mix. These models can support targeted interventions ahead of regulatory cycles.

3. Customize Quality Metrics Based on Facility Profile

Implement facility-type-specific performance targets. For instance, Critical Access Hospitals should not be compared directly with Pediatric or Major Teaching Hospitals due to inherent differences in case severity and volume.

4. Public Transparency as a Performance Lever

Launch a public-facing SIR performance portal where citizens can view their local hospital's yearly infection performance. This can act as both a public health awareness tool and a quality improvement pressure mechanism.

5. Expand Support for Low-Volume Hospitals with High SIR Volatility

Many smaller facilities (e.g., those with one infection but predicted rate close to zero) are unfairly penalized due to statistical instability. Provide statistical smoothing methods or alternate indicators to guide improvement in these settings.

Unexpected Outcomes

1. Improved SIR Despite Rising Infection Counts and Procedure Volumes

Observation: The 2023 average SIR declined to 0.80 despite increases in both procedures (+0.8%) and reported infections (+0.8%).

Explanation: This may indicate better-than-expected improvements in infection control relative to procedure complexity or reflects more accurate prediction modeling for infections.

2. High SIRs in Facilities with Very Few Infections

Observation: Facilities such as George L. Mee Hospital (SIR = 6.67) and Children's Hospital of Orange County (SIR = 7.69) reported only one infection each.

Explanation: These outliers likely occur due to extremely low predicted infection values. A single infection can inflate the SIR significantly when the baseline is near zero. This highlights a limitation in how SIR behaves in low-volume contexts and suggests a need for alternate metrics or threshold-based adjustments.

3. Top Counties for High SIRs Did Not Have High Infection Volumes

Observation: Yuba and Inyo had the highest SIRs but relatively low absolute infection counts.

Explanation: SIR is sensitive to predicted infection rates. When these predicted rates are low (as seen in rural or small hospitals), even a few infections drastically affect the ratio. Policymakers should consider SIR volatility when applying penalties or incentives.

Conclusion

Key Learnings

1. Stable Yet Rising Risk Profile: Despite an average SIR of 0.80 in 2023—indicating infections remain below the safety threshold, the increase in both reported and predicted infections year-over-year suggests a subtle escalation in healthcare-associated infection (HAI) risks.

2. Facility-Level Disparities: Specific facilities like George L. Mee Hospital and Children's Hospital of Orange County report exceptionally high SIRs (6.67 and 7.69, respectively), despite low infection counts. These highlights possible gaps in infection control protocols or small denominator bias due to fewer procedures.

3. Geographic Variability in Risk: Counties like Yuba and Inyo show disproportionately high SIRs with limited facility counts, suggesting either localized outbreaks, data irregularities, or poor infection control practices in remote areas.

4. Procedure-Driven Vulnerabilities: Certain surgeries especially heart transplants, abdominal aortic aneurysm repair, and pacemaker procedures consistently yield higher average SIRs. Colon surgery also tops the list for absolute infection counts, flagging it as a high-risk operative area.

5. Benchmark Distribution and Compliance: A large number of facilities (1,104) perform worse than benchmark SIR levels, while only 6 matched benchmarks, indicating widespread underperformance and uneven implementation of infection control strategies.

Limitations

- 1. Incomplete Date Data:** The absence of precise timestamped data limited the ability to perform granular time-series analysis (e.g., seasonal trends, monthly fluctuations), restricting insights to annual summaries.
- 2. Limited Contextual Variables:** Lack of patient-level variables (age, comorbidities, length of stay) and facility operational data (staffing levels, ICU capacity) limited the ability to control for confounding factors in SIR variability.
- 3. Outliers with Small Counts:** Facilities with extremely high SIRs but very low infection or procedure counts may distort interpretation due to statistical noise from small samples.
- 4. No Confirmed Causal Links:** While correlations between facility type, county, and SIR exist, the analysis cannot establish causality due to the observational nature of the dataset.
- 5. Current Limitations:** Annual-only timestamps limit seasonal analysis. - Lack of clinical metadata (e.g., patient age, length of stay). - SIR distortion in low-volume facilities.

Future Research

- 1. Integrate Temporal Data:** Collect and analyze month-by-month infection trends to identify potential seasonality, outbreak clusters, or temporal policy effects.
- 2. Augment with Clinical Metadata:** Incorporate patient-level health records (age, diagnosis, comorbidities) and hospital operational metrics (nurse-patient ratio, hygiene audit scores) to improve model precision and explainability.
- 3. AI Model Enhancement:** Use gradient boosting or neural networks to predict high-risk combinations of procedures and facility types. This would allow hospitals to allocate infection control resources more intelligently.
- 4. Longitudinal Performance Tracking:** Develop a dynamic dashboard that tracks the SIR performance of each facility over multiple years to detect persistent underperformance or notable improvements.

5. Incorporate Genomic Surveillance Data: For facilities or counties with persistently high SIRs, future studies can explore the use of pathogen genomics to determine if certain bacterial strains or resistance profiles contribute to repeated infection patterns.

6. Integrate clinical and genomic data for better causal inference: Monthly infection tracking to detect seasonal patterns. - Implement deep learning for patient-level infection risk scoring.

References

CDC (2024). Surgical Site Infection (SSI) Prevention Guideline. [online] Infection Control. Available at: <https://www.cdc.gov/infection-control/hcp/surgical-site-infection/index.html>.

www.cdc.gov. (2023). 2015 Rebaseline | NHSN | CDC. [online] Available at: <https://www.cdc.gov/nhsn/2015rebaseline/index.html>.