ICU Patient Report

By: Taylen Inthisane

Introduction

Regional Medical Center is a 400-bed medical center that serves a diverse population. The hospital currently has 36 beds allocated to intensive care unit (ICU) patients. ICU patients tend to require more monitoring and care than the average patient. Regional Medical Center wants to improve patient outcomes while maintaining their costs. Due to the considerable variability in patient characteristics, the ICU Medical Director has commissioned a data analysis to determine whether specific attributes influence patient care and recovery outcomes. To accomplish this task, I will adhere to the CRISP-DM framework, beginning with business understanding and progressing to data understanding, to systematically identify the factors that influence patient recovery.

Section 1: Business Understanding

Section 1.1: Background

Organizational/Context Description: Regional Medical Center (RMC) is a 400-bed medical center serving a diverse metropolitan population. It's 36-bed intensive care unit (ICU) treats the most critically ill patients, and the hospital is focused on improving patient care by leveraging electronic medical records (EMR's) to identify key factors that influence ICU patient recovery. Through this data-driven analysis, RMC aims to generate actionable insights that can guide clinical practice and enhance patient outcomes.

Domain and Context Assessment: Research on ICU admissions highlights ongoing debate about whether patients are being appropriately admitted to intensive care. For example, a University of Michigan study of pneumonia patients over the age of 65 found that one in five were placed in the ICU, raising questions about overuse of scarce ICU beds for patients who might not require that level of care. (Valley & Cooke) This controversy underscores two gaps in current understanding: how to accurately classify patients for ICU versus general admission, and how such classifications affect both resource allocation and patient well-being. While clinical outcomes are often measured, the psychological impact of being labeled as an "ICU patient" remains unexplored, leaving important questions unanswered about the broader consequences of ICU admission practices.

Problem Situation: Staff at Regional Medical Center (RMC) are increasingly concerned about the variability of ICU patient outcomes. Despite extensive clinical expertise, it remains unclear which patient characteristics, treatments or care processes most strongly influence recovery, length of stay or resource utilization. Current approaches rely heavily on individual clinical judgment, which makes it difficult to identify consistent patterns across diverse medical conditions. By systematically analyzing electronic medical records, RMC hopes to uncover common factors that can guide evidence-based improvement in ICU care and provide actionable insights for both care clinicians and hospital leadership.

Data Science Justification: With comprehensive electronic medical records, data science approaches enable systematic analysis of ICU patient characteristics, treatments, and outcomes that are difficult to detect through clinical observation alone. Statistical and machine learning methods can uncover relationships – such as which patient profiles are linked to longer ICU stays, higher resource utilization, or faster recovery – that traditional judgment-based approaches may miss. By moving beyond isolated case experience, data-driven insights can provide clinicians with evidence-based guidance for patient care

while helping administrators allocate resources more efficiently, ultimately improving both patient outcomes and hospital performance.

Stakeholder Identifications:

Primary Stakeholders: The ICU Medical Director, responsible for overseeing clinical protocols and patient outcomes. Their primary focus is on evidence-based insights to guide treatment decisions and elevate care quality. These findings help refine ICU admission criteria, streamline care pathways, and enhance recovery practices. Technical sophistication: High in clinical expertise, moderate in data interpretation.

Secondary Stakeholders: Hospital administration, accountable for operational performance, cost control and resource allocation. They are focused on how ICU efficiency and patient outcomes impact overall hospital performance. Insights from this research would guide decisions on policy, staffing, and budgeting. Technical sophistication: High in operations and finance, limited in data analytics.

Commented [GU1]: Should this be "cost control" and not "cost, control"

Technical Stakeholders: The hospital's data analytics and IT/clinical information teams oversee EMRs, data pipelines, and statistical modeling, with a focus on delivering accurate, secure, and scalable analysis. They directly implement methodologies and maintain the systems that support ongoing analytics. Technical sophistication: High in data science and technology.

Section 1.2: Project Objectives and Success Criteria

Primary Objective: Determine whether statistically significant relationships between ICU patient characteristics, treatment patterns, and outcomes such as length of stay, recovery, and resource utilization. The goal is to provide evidence-based conclusions that address current gaps in understanding and support RMC clinicians in refining care practices while helping administrators make informed operational decisions.

Secondary Objectives: Secondary objectives identifying treatment patterns that most strongly influence patient recovery, as well as those that appear to have limited or negligible effects. Assess whether specific patient demographics, such as age, are significantly correlated with longer ICU stays.

Quantitative Success Criteria: Statistical significance $p<0.05$ must be achieved for all key relationships and practical effect size for health outcomes. Relationships that meet both criteria will be considered validated and contribute actionable insights into clinical or operational decision making. Relationships that do not reach significance are still valuable, as they indicate areas where there are no associations.

Qualitative Success Criteria: The ICU Medical Director demonstrated confidence in the analytical methodology used to understand patient recovery factors, highlighting positive patient responses to care-informed recommendations and valuable contributions that support hospital administration in optimizing ICU resources utilization and cost management.

Impact Assessment: Providing evidence-based recommendations to hospital administration to improve their current clinical procedures that will lead to enhance ICU patient care. Patients will see improved care, decreased time spent in the ICU. While the hospital will receive the benefit of decreased patient mortality, maintained costs while providing better services and improved brand reputation from providing outstanding services.

Timeline and Milestones:

- Weeks 1-4: Data integration and quality assessment.
- Weeks 5-7: EDA plus Statistical analysis and data modeling.
- Weeks 8-9: Data Modeling.
- Weeks 10- 11: Data evaluation and refinement.

- Week 12: Final production
- Week 13: Publication and share with stakeholder.
- Week 14+: Training and roll out of changes.

Section 1.3 Assessment of Responsibility

Legal and Regulatory Considerations: Because this project involves electronic health records, compliance with the Health Insurance Portability and Accountability Act (HIPAA) is essential. The HIPAA privacy rule protects all individually identifiable health information (PHI), including past, present, and future physical or mental health conditions. Our analysis must therefore ensure that the data are either de-identified according to HIPAA's Safe Harbor and Expert Determination standards, or otherwise handled under strict access controls, Re-identification of patients from the dataset is explicitly prohibited. In addition, HIPAA security rule requires appropriate safeguards, technical, administrative and physical to protect the confidentiality and integrity of the data throughout the project lifecycle.

Privacy and data protection: Healthcare data used in this project must comply with HIPAA's de-identification standards, which can be achieved, through (1) a formal determination by a qualified statistician or (2) removal of all specified identifiers. This ensures that individual patients cannot be re-identified from the dataset.

Beyond direct identifiers, we must considered sensitive inferences, for example combination of demographic, diagnoses or treatment patterns, could inadvertently reveal information about a patient's condition or risk profile. To mitigate this, analysis outputs will be aggregated and reviewed to prevent unintended disclosures.

At the collective level, results must be presented in a way that avoid stigmatizing demographic groups of ICU populations, framing insights in a balanced and context-sensitive manner.

From a data security perspective, RMC must maintain a secure database with encryption, strict access, controls and audit trails. Equally important are administrative safeguards, including employee training to prevent misuse of data and compliance breaches.

Bias and Fairness Concerns: Age related differences in patient health care introduce systemic bias in the analysis. For example: older patients often experience longer recovery times or higher complication risks due to biological factors. If not carefully controlled for, the data may incorrectly suggest that certain treatments are ineffective or that clinical care is failing, when in fact outcomes are influenced by age-related vulnerability.

Additionally, representation gaps may exist if certain patient populations (elderly, lower-income, minorities) are underrepresented in the dataset. This could limit the generalizability of finding and results in recommendations that do not equability serve all ICU patients.

Stakeholder Impact Analysis:

Primary Stakeholders: The ICU Medical Director serves as the primary stakeholder, responsible for reviewing the analysis and implementing clinical process changes that improve patient care. These changes will extend to the ICU care team, reshaping daily procedures. While the benefits include more effective, data-informed care, potential drawbacks include training costs and temporary labor shortages during the transition, which could momentarily impact patient care quality.

Secondary Stakeholders: Hospital administration, finance, purchasing, and logistics departments are secondary stakeholders. They will support the implementation by coordinating resource allocation, supplier selection, and departmental workflows to ensure cost-effective, high-quality care. The main

challenges include additional administrative overhead, interdepartmental coordination, and potential delays due to lead times and resource availability.

Vulnerable Population: Special attention must be given to elderly, minority, and economically disadvantaged patients. Analysis and subsequent recommendations must ensure equitable treatment, avoiding disadvantage due to language, comprehension, age, or economic barriers. Machine learning and data-driven insights should be carefully monitored to prevent unintended bias against these groups.

Differential Impact Assessment: The impact of ICU care and data-driven recommendations is not uniform across patient populations. Elderly patients often experience longer recovery times and higher susceptibility to complications due to weaker immune systems. Economically disadvantaged patients may face barriers to accessing preventative care or necessary medications, which can exacerbate illness severity and influence ICU outcomes. Language or comprehension barriers can also limit understanding of care instructions or adherence to treatment plans.

Additionally, clinical context matters: patients with different conditions such as trauma versus acute infections experience vastly different care pathways and recovery trajectories. The analysis and subsequent recommendations must account for these differences to ensure equitable treatment, prevent inadvertent bias, and provide insights that are actionable across diverse patient populations.

Mitigation Strategies: Technical safeguards include applying HIPAA safe harbor or expert determination methods to remove or mask patient identifiers. Process oversight includes requiring ICU clinicians and data governance staff to review models, outputs and recommendations before implementation.

Section 1.4 Data Science Goals and Success Criteria

Technical Problem Framing: The analysis will be framed as a regression problem to quantify relationships between patient characteristics, treatment patterns, and care processes to length of stay, resource utilization and patient recovery.

Data Science Objective: Primary data science goal is to establish statistically significant correlation between (patient characteristics, treatment patterns and care processes) and (length of stay, resource utilization and patient recovery) with a 95% confidence interval, secondary goal is to isolate additional demographics affecting patient care to better utilize resources.

Technical Success Criteria: Detect relationships between patient characteristics, treatment patterns, and care processes and ICU outcomes (length of stay, resource utilization, and patient recovery) with statistical significance ($p<0.05$) and effect size large enough to support practical interpretation. Models must explain meaningful variance after controlling confounders and provide confidence intervals suitable for guiding actionable clinical and operational decisions.

Analytics Approach Overview: The analysis will begin with descriptive statistics to summarize patient characteristics, treatment patterns, and ICU outcomes, followed by visual exploration using scatter plots, pair plots, and residual plots to identify potential trends and relationships. Regression analysis will then be used to quantify the relationships between variables, isolating the effects of individual patient and treatment factors on outcomes such as length of stay, resource utilization, and patient recovery. Feature selection will combine domain knowledge from clinical expertise with data-driven insights to ensure meaningful and interpretable models.

Objective-Technical Mapping: Statistically significant regression coefficients -> credible evidence for procedural changes -> findings communicated to ICU Medical Director and Hospital adminstration ->

updated clinical procedures are implemented -> ICU resource allocation improved -> Potential increase in ICU patient recovery and overall wellbeing.

Constraints and Assumptions: Limited age data due to patient de-identification may affect resource allocation assumptions, particularly for elderly patients who typically require more intensive care. Inconsistent medication administration across patients introduces variability in outcomes that may not reflect clinical efficacy. Socioeconomic factors influencing health, such as nutrition and access to care, are not captured and remain outside the scope of the analysis. Assume observed ICU outcomes reasonably reflect patient recovery and resource needs despite these data limitations.

Section 1.5 Project Plan

Phase-Specific Planning: Data preparation phase will involve aggregating multiple data frames by merging on unique patient identifiers to create a unified dataset for analysis. All variables being tested will be standardized, and missing values will be handled, dropped or imputed based on the context of the treatment data. One-hot encoding may be applied to categorical variables, such as treatments, to explore correlations with patient recovery.

Modeling: Predictive modeling will begin with standard OLS linear regression to quantify relationships between patients' characteristics, treatment patterns and ICU outcomes. To address multicollinearity and improve feature selection, Lasso and Ridge regressions will also be applied. These regularized models will help identify the most influential variables while ensuring stable and interpretable coefficient estimates for actionable insights.

Evaluation: Model performance will be assessed from both operational (business) and clinical (medical) perspectives. From the business perspective, feasibility and cost effectiveness of implementing proposed processes will be evaluated, leveraging financial and operational knowledge. From the medical perspective, priority will be given to interventions that improve patient outcomes and overall ICU care quality. These perspectives will be weighted 40% business and 60% medical to select the model that best balances resource efficiency with patient-centered care, ensuring alignment with the hospital's overall objectives.

Deployment: Deployment will begin with a meeting in the first week between the ICU Medical Director and project leads to review analysis findings and discuss strategies for integrating updated clinical procedures. In the second week, a panel of hospital administrators and ICU leadership will convene to define and agree on measurable benchmarks, such as reducing clinical response time for administering corrective medications where applicable. Once benchmarks are established, the subsequent months will focus on training clinicians to ensure proficiency in the new procedures. After implementation, the ICU Medical Director and hospital leadership will hold monthly review meetings to monitor progress against benchmarks, evaluate outcomes, and make necessary adjustments.

Timeline and Milestones:

- Week 1-3: Review business objectives, meet with stakeholders to clarify expected outcomes, and finalize the project plan.
- Weeks 4-5: Review the data, consolidate ICU datasets into a master dataset for initial analysis and report the findings.
- Week 6-11: Initiate data preparation and modeling by standardizing variables, addressing missing values, and engineering relevant features. Apply regression models such as OLS, Lasso, and Ridge to examine relationships between patient characteristics, treatments and outcomes. Model performance will be evaluated from both medical (60%) and business (40%) perspective to ensure balanced insights.

- Week 12-14: Refine the analyses, finalize reports and visualizations, and prepare presentations for stakeholders. Share results with ICU leadership and hospital administration, highlighting actionable recommendations and benchmark guidance.

Risk Identification and Contingencies:

Risk: Unmeasured patient behaviors and health factors such as smoking, exercise, dietary issues may influence outcomes and introduce confounding variables.

Contingency: Use a conservative approach to standardize patient health status assumptions, treating unknown factors as higher risk. Where possible, backfill missing or incomplete data to enhance assessment accuracy.

Risk: Stakeholder availability or scheduling conflicts may delay feedback and implementation.

Contingency: Incorporate buffer periods into the timeline, deliver preliminary reports ahead of schedule, and leverage asynchronous feedback tools to maintain project momentum.

Risk: Missing or inconsistent data may hinder the accurate modeling of recovery outcomes.

Contingency: Use imputation techniques where appropriate or conduct sensitivity analyses to evaluate the impact of missing data on model outcomes.

Section 2: Data Understanding

2.1 Data Inventory:

Primary Dataset Description: Primary dataset is the icustays.csv(22KB, CSV format) which contains 140 ICU stays for 100 unique patients from Medical Information Mart of Intensive Care (MIMIC). The dataset includes the following fields: patient identifier, hospital admission ID, stay ID, first care unit, last care unit, in-time, out-time, and length of stay. Each row represents a single ICU stay for a patient. Additional important datasets include: admissions.csv(47KB, CSV Format) which contains information regarding a patient's admission and patients.csv(3,313KB, CSV format) which contains information regarding the patients' demographics.

Additional Data Sources: No supplementary datasets were used in my analysis, as the primary dataset—spanning 14 years of ICU electronic medical records from a diverse metropolitan population—offered ample temporal depth and demographic variety. This richness provided sufficient coverage to identify seasonal trends and support meaningful insights, making the inclusion of external sources unnecessary for the scope of this study.

Data Collection Methodology: The MIMIC ICU dataset comprises records from patients admitted to the ICU or emergency department at Beth Israel Deaconess Medical Center (BIDMC) between 2008 and 2022. Data was captured through the hospital's electronic health record (EHR) systems and retrospectively aggregated for research purposes, meaning entries were compiled after the fact rather than updated in real time within a patient's master record. The dataset includes the full population of ICU admissions during this 14-year period, with no sampling and every recorded stay is represented. However, slight changes in EHR systems or documentation practices over time may have introduced inconsistencies in data formatting or recording frequency. Additionally, patients with incomplete or corrupted medical records are excluded, which presents a limitation in the dataset's comprehensiveness.

2.2 Data Description

Comprehensive Data Dictionary: Documenting 18 of the 132 different variables recorded.

Chart time (datetime, any date from 2008-2022): date a chart was created.

Stored time (datetime, any date from 2008-2022): The date the chart was stored.

In_time (datetime + hour, any date from 2008-2022 with valid exact time): Date and time person was administered in the ICU.

Out_time (datetime + hour, any date from 2008-2022 with valid exact time): Date and time person was discharged from the ICU.

Length of_Stay (LOS) (float64 , 0-30+) the length of stay for an ICU patient.

Creation_minutes (float 64, 1-9040) minutes it takes to create medicine.

Storage_minutes (float64, -9054 - 3054) Time it takes to store create medicine, could be created and stored before.

Gender (object, Male, Female) Gender of patient.

Anchor_age (int64, 21-90+) Age of the patient.

Death_Status (object, Alive, Dead): If the patient is alive or dead.

Race (object, White, Black, etc.) What race is the patient.

Marital_status(object, Married, Single, Divorced, Widowed) Patients marital status.

Language (object, English, Other) What language the patient speaks.

Insurance (object, Other, Medicare, Medicaid) what type of insurance the patient has.

Admission_type (object, Surgical Same day, etc.) Reason why a patient being admitted.

Admission Location (object, physician referral, etc.) where they were admitted from.

Discharge location (object, home, hospice, etc.) where the patient was discharged to.

Treatment Category (object, Antibiotic, Medications, etc.) what type of treatment the patient received.

Dataset Structure and Relationships: The current structure of the dataset is multiple related tables. The data was denormalized and is in a star schema setup. The reason a star schema is adapted is because not all data is applicable for a department. The pharmacy might not need to know the patient's length of stay but they need to know the patient's body weight. The star schema reduces unnecessary data while having the ability to create a patient's complete profile through patientIDs. Each line represents a distinct medical record entry. Most patientID is recorded multiple times with each occurrence corresponding to a unique record. The hierarchy is Patient->provider/hospital admin -> stayid -> documented times -> events.

Temporal Coverage and Granularity: The data covers a 14-year period (2008-2022) representing the full population of ICU patients at RMC. Documentation is ongoing, as the hospital strives to maintain up-to-date records for every patient. Given the retrospective aggregation over such a long span, some procedural changes have occurred. For example, the "value" column contains a mix of information, and certain timestamps were not consistently recorded. However, since RMC operates in a single geographic location, time zone inconsistencies are not a concern.

Data Volume and Scale: Entire dataset decompressed is 122.8 MB with a split of 74MB coming from ICU data and 48.8 MB coming from Hospital data. By joining these two data sets together, creates a dataset that can be used for data analysis. These datasets both share a primary key of subject_id and by joining them together, a more cohesive picture of the ICU patient and hospital care can be created.

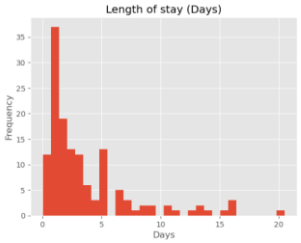
Categorical variable encoding : Out of the categorical variables majority of them have the exact description of the item. However, there are two categorical variables that stick out Isopenbag and Continueinnextdepartment. Both variables are one hot encoded, where a value of 1 indicates "yes" and 0 indicates "no". For examples, in the case of Isopenbag, a value of 1 means the patient has a Bogota bag, while a value of 0 means they do not.

Commented [GU2]: Is this one word or spelling error?

Commented [TI3R2]: One word.. It's a variable

Commented [GU4R2]: Thanks!

2.3 Data exploration report



Univariate Analysis: The chart creation timestamps revealed an interesting discrepancy: the top 10 creation dates did not align with the 10 storage dates. For instance, the most frequent chart creation date was 2146-06-26 at 8:00 am yet this date did not appear among the top storage dates. Instead, the most frequent chart storage date was 2137-02-27 at 12:04 pm. While I did not expect the dates to match exactly, it is surprising that the most common creation and storage dates are not even on the same day.

The distribution of length of stay is right-skewed, with most patients recovering within five days and a smaller group requiring longer stays. Notably, the average ICU stay is 3.679 days, and 75% of patients are discharged by day 4.9.

The monthly ICU intake does not follow a normal distribution. July shows a noticeable peak in visits, likely due to increased outdoor activity during warmer weather. Intake then declines before rising again as colder temperature set in and winter approaches.

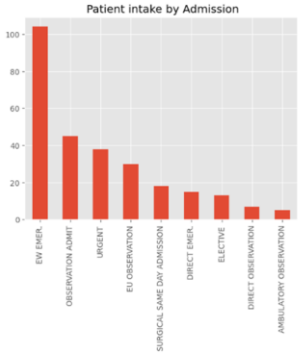
The ICU discharge rate reveals some intriguing seasonal trends. July continues to show high activity, consistent with overall patient intake. In contrast, winter months see a surge in admissions, likely due to seasonal health issues. Notably, March stands out with an unusually sharp spike in discharges, suggesting a significant post-winter recovery period.

Medicine creation and storage timestamps were examined, but since both actions typically occurred immediately and simultaneously, they offered little variation. As a result, these variables did not yield meaningful insights or trends.

The gender distribution of ICU admissions is nearly balanced, with 57% male and 43% female patients.

The age distribution of ICU patients is left-skewed, indicating that older individuals are more commonly admitted. The average age of ICU patients is 61.75 years.

Patient Admissions: The majority of ICU admissions originate from the emergency ward, followed by admissions through observation and then urgent referrals.



The majority of patients admitted to the ICU were white, with smaller proportions represented across other racial groups.

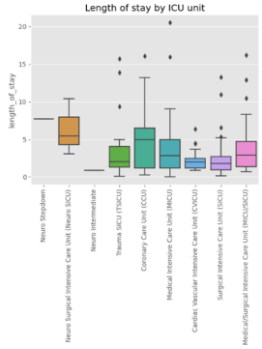
Marital status: Majority of the patients that were admitted into the ICU were single.

Languages: Most ICU patients were English speaking.

Majority of the patients had private or individual insurance plans, while just over 100 were covered by Medicare and approximately 20 were enrolled in Medicaid.

Bivariate Relationships: When comparing length of stay with both chart creation and storage times, no discernible patterns emerged. The data appears to be randomly distributed, offering little insight into any meaningful trends.

The length of stay data by ICU unit revealed some compelling differences, with the Coronary Care Unit showing significantly greater variation compared to other units.



Analysis of length of stay in relation to medicine creation and storage times revealed no significant trends, suggesting these variables have little impact on patient duration in the ICU. Similarly, no meaningful differences were observed between genders; the violin plot indicates that both male and female patients tend to stay within the same dense range of 0 to 5 days. Age also showed no clear correlation with length of stay, offering limited insight into age-related patterns. However, some variation was noted based on admission type particularly among patients admitted through the emergency ward and direct emergency routes which may warrant further investigation.

Medicare patients had the longest average length of stay among insurance types, which aligns with expectations given that the average ICU patient age is 61. This age group is more likely to be covered by Medicare, contributing to the broader variation in stay duration.

There were no significant findings regarding length of stay by race or language. While white patients showed the greatest variation, they also represented the largest demographic group, and English was the dominant language limiting the depth of insights from these variables.

Marital status revealed an interesting pattern: married patients had the second-shortest mean length of stay, yet their box-and-whisker plot was more spread out compared to other marital groups, which showed tighter distributions.

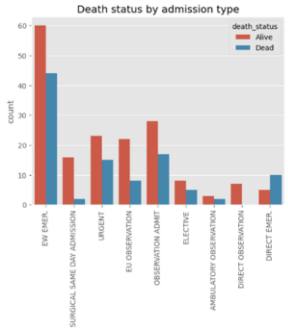
Procedure type also offered a notable observation. Patients undergoing respiratory procedures had stays ranging from 5 to 15 days, suggesting that enhancements in respiratory care could potentially reduce ICU duration.

Lastly, there were no dramatic differences in death status by gender or age. However, it's worth noting that deceased patients tended to be older on average than those who survived.

Death status by race revealed an interesting trend: Black/African American patients had a disproportionately higher death count compared to other racial groups. This suggests potential disparities that may warrant further investigation.by

An interesting pattern emerged when examining death rates by insurance type: Medicaid had a higher proportion of deceased patients compared to those who survived, especially when contrasted with other insurance categories.

Death rates by admission type revealed a notable trend: patients admitted through direct emergencies had a higher mortality count compared to those who survived.



Death rates by first care unit highlight the Medical Intensive Care Unit as a potential area for further investigation, as it recorded more patient deaths than survivals.

Home discharges showed a surprisingly high death count relative to the number of surviving patients, indicating a potential area for further review or concern.

Temporal Patterns and Trends: Temporal patterns and trends aren't particularly relevant to this project due to its scope. ICU procedures tend to remain consistent regardless of the season. For example, if someone requires surgery for a bullet wound, the medical care provided will be the same whether it's winter or summer.

Unexpected Discoveries: One of the more surprising discoveries was that patient age isn't a strong determining factor in ICU outcomes. I initially expected older patients to fare worse due to weaker immune systems, but the data didn't support that assumption. Another interesting finding was that patients undergoing respiratory procedures tend to have longer ICU stays than those with other conditions. I had assumed cardiac procedures would lead to the longest stays, given their complexity and delicacy. Lastly, a striking insight was the high mortality rate among Medicaid patients in the ICU. This may be linked to financial barriers that delay care until it's absolutely critical. It was particularly noteworthy that more Medicaid patients died than survived during their ICU stay.

Organizational-Relevant Segmentation: A particularly relevant segment is the Medical Intensive Care Unit, which stands out as a potential focus area for the ICU director. Given the high number of patient deaths in this department, it presents an opportunity to explore targeted improvements in patient care and outcomes.

Hypothesis Generation:

Hypothesis: Patients covered by Medicare tend to have longer hospital stays and require more intensive care compared to those with other types of insurance.

Hypothesis: Patients on Medicaid tend to exhibit a higher mortality rate in the ICU, despite having the shortest average length of stay. This may suggest that financial constraints lead some patients to delay seeking care until absolutely necessary, potentially impacting outcomes.

Hypothesis: Discharging patients to home without home health care support poses a higher risk of mortality, likely due to insufficient follow-up care and medical oversight.

2.4 Data Quality assessment

Completeness Analysis: When evaluating the completeness of the data, one significant missing characteristic that impacted my analysis was age. Due to HIPAA privacy regulations, data on minors cannot be recorded or included in the MIMIC dataset, resulting in their underrepresentation. This poses a challenge, as ICU care spans all age groups. Another underrepresented characteristic is language. In the dataset, language is categorized simply as "English" or "Other," which overlooks widely spoken languages like Spanish. This broad grouping limits the depth of analysis and introduces an additional barrier to understanding patient outcomes. For non-English speakers, miscommunication with healthcare providers may occur, potentially leading to misunderstandings and compromised care.

Accuracy Assessment: Potential errors emerged during the grouping of procedure and diagnosis codes due to inconsistencies in their formatting. Some codes were two digits, others three or four, which made standardization challenging. While categorizing procedure types, I encountered an unusually high number of deaths attributed to male genital procedures around 60 which seemed implausible given the context of ICU care. This suggested a misclassification issue. Similarly, I found over 800 deaths listed under "unknown diagnoses," which was clearly inaccurate considering the dataset only included 140 patients. These discrepancies highlight the importance of careful code handling to ensure valid analysis.

Consistency Evaluation: One of the key issues I encountered with the input, output, and procedure events was the handling of time. The start, end, and store times were treated as generic objects rather than properly formatted datetime values, which complicated temporal analysis. Another consistency challenge involved procedures and diagnoses since a single patient can have multiple entries for each, it became difficult to accurately assess length of stay and mortality based on those categories.

Timeliness and Currency: The most recent data available is from 2021, creating a roughly four-year gap from the present. However, the dataset remains highly relevant due to its extensive 14-year span, which not only provides a robust historical view of hospital admissions and discharges but also includes the COVID-19 pandemic period. This combination offers a valuable opportunity to analyze patient care both before and during a global health crisis, making the dataset especially useful for identifying trends and evaluating healthcare responses over time.

Relevance and Coverage: After reviewing the dataset, it's clear that it doesn't represent the full population of ICU patients. Due to HIPAA privacy regulations, data on minors is excluded, resulting in a significant portion of ICU patients being left out. This gap limits the dataset's ability to fully reflect the diverse age range of individuals who require intensive care.

Data Quality Impact Assessment: The inability to sort by procedure and diagnosis groups significantly limits my ability to make informed recommendations to the ICU medical director. Inconsistencies in the data prevent meaningful insights from being drawn from these categories. For instance, there may have been more effective procedures for patients with cardiovascular diagnoses, but without reliable grouping, such patterns are difficult to identify. As a result, the analysis must rely more heavily on patient characteristics such as age, insurance type, and admission method to assess recovery outcomes and length of stay.

Conclusion: During the initial stages of the CRISP-DM process, business objectives were clearly established, guiding the selection of key variables such as length of stay, mortality status, and treatment category for early exploration. Preliminary analysis uncovered notable trends, including a surge in ICU admissions during July, a high volume of patients entering through the emergency ward, and significant variability in individual ICU stays. This surface-level exploration lays the groundwork for more advanced data preparation and modeling, revealing patterns and characteristics that merit deeper investigation. These insights will shape the next phases of analysis, ensuring a focused approach that supports improved ICU resource planning and enhances patient care outcomes.

Work Cited

Valley, T., & Cooke, C. (n.d.). *Who needs to be in an ICU? It's hard for doctors to tell*. https://ihpi.umich.edu/news/who-needs-be-icu-it%E2%80%99s-hard-doctors-tell.