Dashboard features information

**📌 KPI Cards in Insights Tab**

**🧮 1. 👥 Total Patients**

**What it shows:**

The total number of patient records currently included in the dashboard (from the dataset readmission\_test\_data.csv).

**How it’s calculated:**

python

CopyEdit

total\_patients = len(data)

**What you can say:**

“This shows the size of the patient population the model was applied to. It helps the admin know how many patients are represented in the insights and predictions.”

**📉 2. 📊 Readmission Rate**

**What it shows:**

The percentage of patients in the dataset who were readmitted within 30 days.

**How it’s calculated:**

python

CopyEdit

readmit\_rate = data['READMISSION\_30'].mean() \* 100

Here, READMISSION\_30 is a binary variable (1 = readmitted, 0 = not readmitted). Taking the mean gives the proportion of patients with a value of 1.

**What you can say:**

“This is the baseline readmission rate across our dataset. It gives admins a sense of how common the issue is across all patients.”

**🛌 3. 🛏️ Avg LOS (Average Length of Stay)**

**What it shows:**

The average number of days patients stayed in the hospital.

**How it’s calculated:**

python

CopyEdit

avg\_los = data['LOS'].mean()

**What you can say:**

“Length of stay is a major factor linked to readmission. If it’s unusually high, it may reflect complicated cases. If it’s very short, patients might be discharged too early — both are useful operational indicators.”

**🧠 How to Present These in Your Video or Q&A**

You might say:

“To start off, the KPIs at the top help hospital administrators quickly understand their dataset. Total patients tells them how many records they’re looking at. Readmission rate sets the baseline of concern. And average LOS helps them assess how long patients are typically staying, which might impact future readmissions.”

**🎬 Dashboard Demo Walkthrough with Example**

**🧑‍⚕️ Scenario:**

"Let's say a hospital administrator wants to check if a patient currently admitted is at risk for readmission. The admin also wants to know which age groups or diagnoses are most at risk, and simulate a patient they’ve just admitted."

**🔍 STEP 1: Predict by Patient**

Go to the **“Predict by Patient”** tab.

🗣️ Say:

"I’ll start by selecting patient index 128. This pulls up their profile:

* Age: 72
* Diagnosis: Chronic Obstructive Pulmonary Disease (COPD)
* LOS: 6 days
* Hemoglobin: 11.2
* BMI: 29
* Smoking status: Current smoker"

"Now I click **Predict Readmission Risk**. The system shows a **risk probability of 0.48**, which is marked as **High Risk**."

🧠 Say:

"This tells the admin that this patient may benefit from extended follow-up or being referred to care coordination. The decision is no longer subjective — it’s backed by patient-specific data."

**📋 STEP 2: High-Risk Patient List**

Now click the **"High-Risk Patients"** tab.

🗣️ Say:

"This view gives the admin a sorted list of all patients with risk ≥ 0.30.  
Here, we see Patient 128 is listed at the top with 0.48 probability.  
Admins can print or export this list to act before discharge."

**📊 STEP 3: Insights Tab**

Switch to the **Insights** tab.

🗣️ Say:

"Let’s explore what factors are linked to higher readmission rates.  
In this Age Group chart, we see patients aged 60–75 have the highest rate."

"This boxplot shows patients with longer stays are more likely to return, and this violin plot confirms lower hemoglobin is associated with readmission."

🧠 Say:

"These visuals support operations planning. For example, the admin might assign more follow-ups to older patients or make policies around discharging patients with borderline hemoglobin."

**📝 STEP 4: Manual Entry Simulation**

Click on **Manual Entry** tab.

🗣️ Say:

"Now let’s simulate a new walk-in patient.  
I’ll enter:

* Age: 75
* BMI: 31
* LOS: 7
* Hemoglobin: 10.8
* Diagnosis: COPD
* Reason: Emergency visit
* Gender: Male
* Smoking: Current smoker
* Race: White"

Click Predict.

"The dashboard shows a probability of **0.57**, clearly marked as **High Risk**. It also warns me that Hemoglobin is **below normal**, helping staff make informed decisions."

**💡 STEP 5: Feature Importance**

Go to **Feature Importance** tab.

🗣️ Say:

"And finally, here’s how the model determines risk.  
As you can see, **DESCRIPTION**, **Hemoglobin**, and **LOS** are the most influential features in prediction."

🧠 Say:

"This gives the admin transparency — they can trust what the model is doing and use that knowledge in care planning."

**✅ Wrap-up:**

"So in this walkthrough, we identified a high-risk patient, simulated a new case, explored trends in readmission, and saw what features drive those predictions.  
This dashboard empowers hospital staff to make **smarter, faster, and safer** discharge decisions."

**🔍 How to interpret it:**

* The **longer the bar**, the more influence that feature has on the model’s prediction.
* The values on the X-axis (Importance) represent **"gain"**, which is how much a feature contributes to improving the decision splits in the trees.
* These features were extracted **after encoding**, so names like cat\_\_DESCRIPTION\_Encounter for problem (procedure) indicate encoded categorical features.

**📌 Example Explanation of Top Features:**

1. **cat\_\_DESCRIPTION\_Encounter for problem (procedure)**
   * This is the **most impactful feature** in predicting readmission risk.
   * Patients admitted for this general "problem" encounter seem to have a strong pattern of readmission, possibly due to unresolved conditions.
2. **cat\_\_REASONDESCRIPTION\_Appendicitis (disorder)**
   * This clinical complaint at admission significantly contributes to readmission—possibly due to post-surgical or follow-up complications.
3. **cat\_\_REASONDESCRIPTION\_Malignant neoplasm of colon (disorder)**
   * Indicates that patients with colon cancer tend to have **higher readmission risks**, perhaps because of treatment complexity or monitoring needs.
4. **cat\_\_GENDER\_M**
   * Although it's the least important among the top 10, it shows that **male gender has some impact** on readmission risk—possibly due to lifestyle or comorbidity differences.

**🧠 Why this matters:**

This feature importance chart helps hospital administrators understand **which patient factors are most critical** when evaluating risk. These insights can guide:

* Discharge planning
* Post-discharge follow-up
* Resource prioritization (e.g., more attention to specific conditions)

**💬 What you can say if asked:**

“Our model used SHAP and gain-based feature importance to identify that the **type of admission** and **underlying diagnosis or clinical reason for hospitalization** are the strongest indicators of 30-day readmission. This aligns with clinical intuition, as certain chronic or acute conditions naturally lead to more complex care needs.”

Let me know if you want a one-slide visual explanation of this too!

**You said:**

can you give me simple explanation how these are considered

**ChatGPT said:**

Absolutely! Here’s a **simple and clear explanation** you can use in your presentation or when your professor asks:

**✅ How these features were considered important:**

1. **Our model (XGBoost)** looks at patient data and learns patterns that help it predict whether someone will be readmitted within 30 days.
2. While training the model, it automatically finds **which features (patient details)** are most useful in making accurate predictions.
3. It assigns **higher importance** to features that help the model make better decisions more often.

**💡 Simple example:**

Let’s say we are trying to guess if a patient will come back to the hospital.

* If the model sees that patients admitted with "Encounter for problem (procedure)" **often get readmitted**, it realizes this feature is very useful.
* So it gives **high importance** to that feature.

On the other hand, something like **gender** may help a little but not as much. So it gets **less importance**.

**📊 Why the chart shows feature names like cat\_\_DESCRIPTION\_XYZ:**

* The model converts categories (like diagnoses) into machine-friendly format using **one-hot encoding**.
* That’s why you see names like:
  + cat\_\_DESCRIPTION\_Encounter for problem (procedure)
  + It means the diagnosis was “Encounter for problem (procedure)” and this had strong influence on the model.

**🧠 In one line:**

These features are considered important **because they most effectively split the patients into high-risk and low-risk groups** during model learning.

Let me know if you want this added as a slide or spoken point for your presentation.

**You said:**

which dataset used to consider this features are this from conditions or procedures

**ChatGPT said:**

Great question! Here's the **exact origin** of these features:

**✅ Which dataset are these features from?**

The top features shown in your **feature importance chart** come from the **conditions** and **encounters** (admissions) datasets in your Synthea-generated EHR data.

Here’s the breakdown:

| **Feature in Chart** | **Comes From** | **Description** |
| --- | --- | --- |
| DESCRIPTION | **Conditions** dataset | The primary medical condition or diagnosis (e.g., "Encounter for problem (procedure)") |
| REASONDESCRIPTION | **Encounters** dataset | The reason or complaint for admission (e.g., "Non-small cell carcinoma of lung") |
| GENDER | **Demographics / Patients** | Patient's gender |
| LOS (Length of Stay) | **Encounters / Admissions** | Total number of days a patient stayed in the hospital |

**💡 In summary:**

* Features like cat\_\_DESCRIPTION\_XYZ → come from **conditions** data.
* Features like cat\_\_REASONDESCRIPTION\_XYZ → come from **encounters** or **admissions** data.
* Together, they form the **input variables** used in the model after preprocessing.

Would you like a small diagram or table slide for your presentation summarizing this?

**1. Train-Test Split**

When training the model, you likely split the full dataset into:

* **Training set** (e.g., 80% = 5,188 patients)
* **Test set** (e.g., 20% = 1,297 patients)

✅ That **1,297 is the test set**, used to **evaluate the model** after training.

Top of Form

**How those high-risk patients are identified:**

1. **Your trained XGBoost model** outputs a **predicted probability** for each patient — a number between 0.0 and 1.0.
2. This number represents the model’s **confidence that the patient will be readmitted**.
3. You set a **threshold** of 0.30:
   * If predicted probability ≥ 0.30 → **High Risk**
   * If < 0.30 → Low risk

**📊 How it's implemented in code:**

Inside the dashboard:

python

CopyEdit

X\_all = preprocessor.transform(data[features])

data['Risk\_Score'] = model.predict\_proba(X\_all)[:, 1]

high\_risk\_df = data[data['Risk\_Score'] >= 0.30]

* model.predict\_proba(X\_all)[:, 1] gives the probability of readmission for every patient.
* Patients with score ≥ 0.30 are added to the high\_risk\_df.

This is what’s shown in your **“High-Risk Patients” table** in the dashboard.

**🔍 What’s included in the table:**

* AGE
* GENDER
* DESCRIPTION (main condition)
* LOS (Length of Stay)
* Risk\_Score (the model’s readmission probability)

**✅ What to say in presentation:**

"We apply our trained model to all patients and calculate their readmission probability. Those with a predicted risk of **30% or higher** are flagged as high-risk and shown to the user. This helps hospital administrators prioritize care or discharge planning for those most likely to come back."

Would you like to customize the threshold interactively in the dashboard too?

Bottom of Form

**Kept after Data Exploration and Statistical Testing**

When we performed **univariate and bivariate analysis**, we checked how these continuous features differed between readmitted and non-readmitted patients using:

* **Mann–Whitney U tests** for statistical significance
* **Box plots** to visually compare distributions

**Hemoglobin** showed a statistically significant difference in readmission outcomes — meaning lower or abnormal levels were linked to higher risk.

| **Feature** | **Test Used** | **Why?** |
| --- | --- | --- |
| **AGE** | Mann–Whitney U Test | AGE was **not normally distributed**, so we used a non-parametric test. |
| **BMI** | Mann–Whitney U Test | BMI distribution was **skewed**, violating normality assumption. |
| **LOS (Length of Stay)** | Mann–Whitney U Test | LOS had high variance and outliers (typical in hospital data). |
| **Hemoglobin** | Mann–Whitney U Test | Hemoglobin values showed skewness and missing values. |
| **Vitals (e.g., Heart Rate, BP)** | Mann–Whitney U Test | These also lacked normal distribution. |

* **LOS and AGE had statistically significant p-values (p < 0.05)** — meaning they differ notably between readmitted vs not-readmitted patients.
* **Hemoglobin also showed significance.**
* **BMI, Heart Rate, etc.** were **not statistically significant**, but were still included because the **model found them predictive** (based on importance scores).

**What is the Mann–Whitney U Test?**

The **Mann–Whitney U test** is a **non-parametric statistical test** used to compare **two independent groups** when the data **is not normally distributed**.

**📊 In your project:**

We used it to compare **numerical features** like AGE, LOS, and Hemoglobin between:

* Group 0: Patients **not readmitted**
* Group 1: Patients **readmitted within 30 days**

**🧠 What it tells us:**

It checks **whether the distribution of a feature is significantly different** between the two groups.

Unlike a t-test, it:

* Doesn’t assume the data is normally distributed
* Works with skewed data or small samples
* Compares **ranks** instead of actual values

**🔍 Interpretation:**

If the **p-value** from the Mann–Whitney U test is:

* **< 0.05** → There’s a **statistically significant difference** between the two groups.
* **≥ 0.05** → No significant difference.

**🧪 Example from your project:**

* **AGE**: p-value = 0.000  
  → Older age is significantly associated with higher readmission.
* **LOS**: p-value = 0.000  
  → Patients with longer hospital stays are more likely to be readmitted.

**🎯 One-liner for presentation:**

“We used the Mann–Whitney U test to identify which numerical features — like Age and Length of Stay — differed significantly between readmitted and non-readmitted patients. This helped us select meaningful predictors for our model.”

**Initial Inclusion Based on Clinical Relevance**

We began by selecting features that are:

* Available in **Synthea's structured datasets**
* Clinically known to influence readmission risk

**Examples:**

* **AGE, GENDER, RACE** – demographic risk indicators
* **DESCRIPTION, REASONDESCRIPTION** – diagnosis/procedure details
* **LOS (Length of Stay)** – longer stays often correlate with severity
* **Hemoglobin, BMI** – vital and lab indicators of health condition
* **Smoking\_Status** – behavioral risk factor

**2. 🔍 Statistical Tests for Numerical Features**

We performed **Mann–Whitney U tests** to check if numerical features (like AGE, LOS, Hemoglobin) were **significantly different** between readmitted and non-readmitted groups.

* **Significant (p < 0.05):**
  + AGE ✅
  + LOS ✅
  + Hemoglobin ✅

These were retained because they showed **statistical evidence of influencing readmission.**

**3. 🧠 Model-Based Feature Importance**

After building an initial XGBoost model:

* We extracted **feature importance scores** from the model itself
* Features like **DESCRIPTION, LOS, and REASONDESCRIPTION** came out as highly predictive
* This confirmed that our **clinical + statistical choices were valid**

**4. 🧪 Encoding + Preprocessing**

* Categorical features were **one-hot encoded** (e.g., DESCRIPTION → cat\_\_DESCRIPTION\_XYZ)
* Numerical features were **kept as is** (e.g., LOS, AGE)

Only features compatible with the model pipeline were retained.

**✅ Final Selected Features:**

| **Feature Name** | **Type** | **Source Dataset** |
| --- | --- | --- |
| AGE | Numerical | Patients |
| LOS | Numerical | Encounters |
| Hemoglobin | Numerical | Observations |
| BMI | Numerical | Observations |
| GENDER | Categorical | Patients |
| RACE | Categorical | Patients |
| Smoking\_Status | Categorical | Conditions/Behavior |
| DESCRIPTION | Categorical | Conditions |
| REASONDESCRIPTION | Categorical | Encounters |

**🎯 What to say in your presentation:**

“We selected features by combining clinical relevance, statistical testing, and model-driven importance scores. This approach ensured that our model is both **data-driven and explainable**, and focuses on variables that truly impact patient outcomes.”

Would you like a slide version of this for your deck?