Annotation Guidelines for eICU ICU Stay Dataset

Course: ARI 510 – HW3 Data Annotation
Project: Predicting Prolonged ICU Stay from eICU-CRD

1. Overview of Data Annotation

Data annotation in this project refers to creating consistent, human-defined mappings for categorical ICU data and clearly defining the meaning of numeric and binary features. This ensures that machine learning models treat equivalent real-world concepts consistently, prevents fragmentation of categories, and improves model stability and accuracy.

2. Annotator Job Description

As an annotator, your responsibilities include:

- Reviewing categorical features and ensuring raw values map to correct canonical categories.
- Verifying the meaning of numeric and binary variables.
- Flagging ambiguous or unclear mappings using the Notes column.
- Following the rules in this document to ensure consistency across annotators.

3. Label Set and Feature Definitions

Each categorical feature has a set of approved canonical categories. These are defined in detail in annotations.xlsx.

4. Annotation Rules

1. Always map raw values to the most appropriate canonical category.
2. Treat raw text case-insensitively and ignore leading/trailing whitespace.
3. If a raw value is ambiguous, map it to the fallback category (other or unknown) and explain in Notes.
4. Do not create new categories; escalate unclear cases.
5. Binary numeric variables must use the stated interpretation exactly.

5. Examples

gender

- "F", "female" → female
- "M", "male" → male
- "nb", "non-binary" → other
- "unknown" → unknown

hospitaladmitsource

- "emergency department", "ED" → Emergency Department category
- "operating room" → Operating Room

- "other hospital" → Transfer from external hospital

6. Annotation Interface Instructions

The Excel spreadsheet annotations.xlsx contains two tabs:

- Categorical Features – Parent rows list features; collapsed child rows list each raw category value.
- Numeric Features – Each feature includes description, value range, binary meaning, and notes.

Using the Categorical Sheet

6. Locate a feature in the feature_name column.
7. Click the + symbol to expand and view category-level details.
8. Review category descriptions and mappings.
9. Use Notes to document concerns or inconsistencies.

7. How Annotations Are Used in the Model

The Python annotation logic normalizes raw categorical values into canonical forms using ANNOTATION_CONFIG. These categories are then one-hot encoded using a fixed ordering to ensure consistent model inputs. Numeric features are imputed and scaled. Accurate annotation ensures reproducible model behavior and improves generalization.

8. Contact Information

- jprantza@umich.edu
- jonto@umich.edu