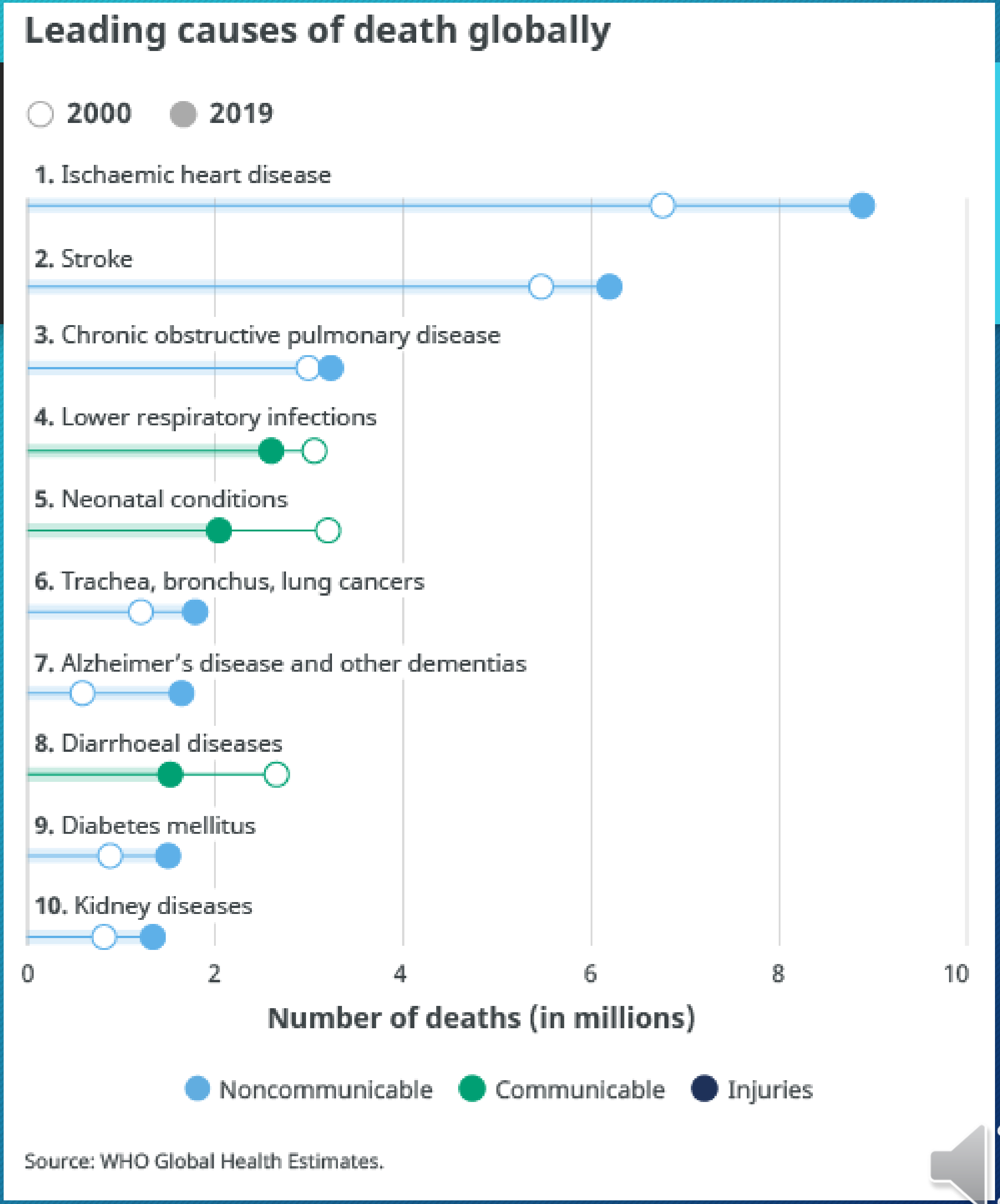
Predicting Stroke Risk

Using individual patient data

Scott Breitbach
22-July-2022

Stroke Statistics

- Worldwide
  - Second leading cause of death
  - 11% of deaths (15 million people)
  - 1 of 6 deaths related to cardiovascular disease
- United States
  - ~800,000 annually
  - Stroke every 40 seconds
  - Death every 3.5 minutes



Leading causes of death globally

Source: WHO Global Health Estimates.

https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death

Stroke Statistics (cont.)

- Of ~15 million annual stroke instances:
  - 1/3 result in death
  - 1/3 recover
  - 1/3 are left disabled
- A leading cause of long-term disability

- Known risk factors:
  - Cardiovascular/Health:
    - High blood pressure
    - High cholesterol
    - Obesity / diabetes
    - Age
  - Other:
    - Race
    - Where you live

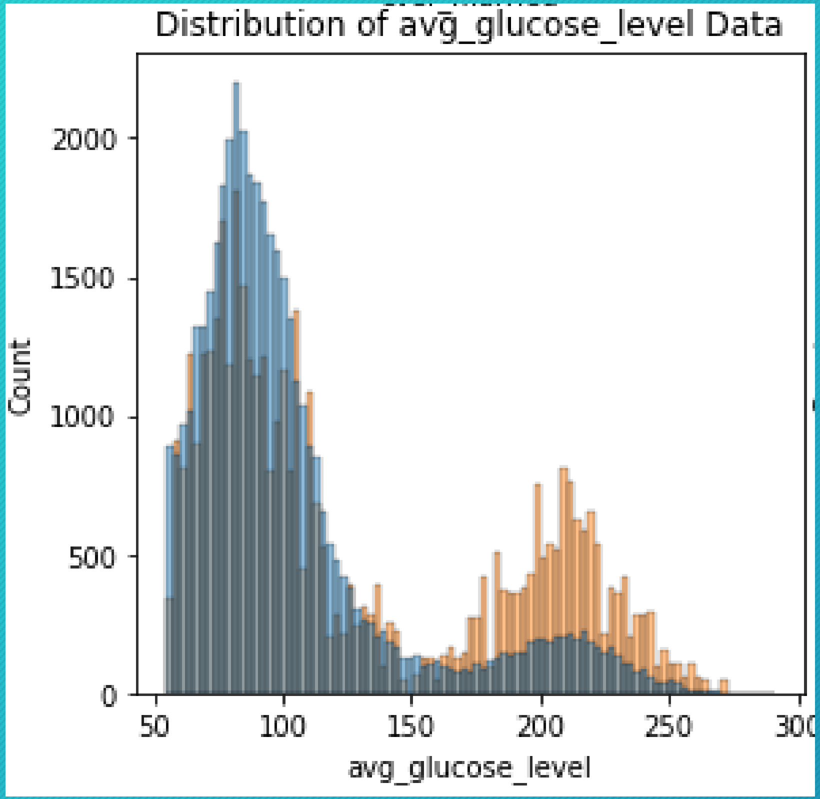
Stroke Dataset: 10 Features

- 5 Categorical:
  - 2 multiclass
    - `work_type`, `smoking_status`
  - 2 binary
    - `ever_married`, `Residence_type`
    - Converted to 1/0
  - 1 multiclass converted to binary
    - `gender` (removed ‘Other’)
    - Converted to 1/0

- 5 Numeric:
  - 3 continuous
    - `age`, `avg_glucose_level`, `bmi`
  - 2 discrete, binary
    - `hypertension`, `heart_disease`
- Target Variable:
  - Binary (1/0)
    - stroke / no stroke
  - Extremely imbalanced
    - Stroke ~2% of dataset

Distributions

- Glucose is bimodal



- Children at lower risk
  - ~6,000 ‘children’ in dataset
- Impacts multiple variables:
  - `age`
  - `hypertension`
  - `heart disease`
  - `ever_married`
  - `work_type`
  - Possibly `smoking_status` (unknown)

Data Preprocessing

- Imputing Null values
  - `smoking_status` 30% Null
    - Null → ‘Other’
  - `bmi` ~3% Null
    - Tried Logistic Regression
    - Landed on median
- Encoding
  - Binary features → 1 / 0
  - Multiclass → one-hot encoded

- Transformation
  - Box-Cox – age, bmi, & glucose
  - Scaled all features -1 to 1
- Balancing (~2% stroke)
  - Oversample stroke
  - Oversample using SMOTE
  - Oversample / Undersample
    - Oversample to 10% of majority using SMOTE
    - Undersample majority so stroke is 50% of majority
  - Leave imbalanced and use weights

Model Selection / Evaluation

Metrics:

- Accuracy
  - 98%, predicting no strokes
- Recall
- Matthews Correlation Coefficient
- Area Under Curve (Receiver Operating Characteristic)

Hyperparameter Tuning:

- Grid Search CV
  - Random Search CV
- Scoring with multiple metrics
- Voting Classifier (with weighting)

Conclusions

- Huge imbalance, not a perfect model
- Need to find a balance in the results
- More features could be helpful
- Possible inherent bias in the data
  - Could be high risk but haven’t had a stroke *yet*

Implementation

- Allow people a level of control over their personal healthcare
- A healthcare app:
  - Answer health questionnaire
  - Store health metrics (weight, blood pressure, health screening results, etc.)
  - Link activity apps (pulse, steps, etc.)
  - Predict risk for stroke, heart disease, and others
  - Provide personalized suggestions for lowering risk
  - A tool to discuss with your primary care physician