Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission

Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M. and Elhadad, N., 2015

Presented By:

Nikhil Verma
Deepkamal Gill

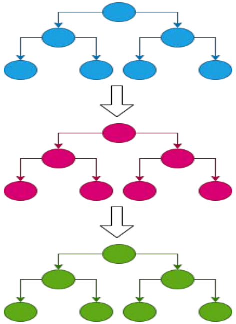
Agenda

- Introduction & Background
- Intelligible Models
- Case Study 1: Pneumonia Risk
- Case Study 2: 30-day Readmission
- Discussion
- Strengths & Limitations

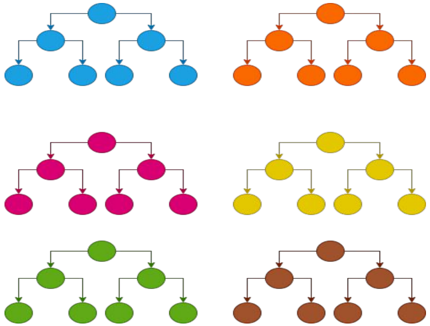
Introduction

- Until recently, humans had a monopoly on the agency in the society
- The reasons for a decision often matter
- Over the past years, rapid progress in ML has led to deployment of automatic decision processes
- Model accuracy and intelligibility generally have a trade-off
- Being able to understand, validate, edit and trust models is critical in healthcare

- **Accurate**

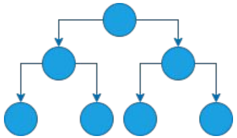


Boosting



Bagging - RF

- **Interpretable**



  - Decision tree
  - Naive Bayes
  - Logistic Regression

Intelligibility: Interpretability by Humans

- Three different ways we can think about intelligibility of model [1]:-

Local vs Global

Local explanation focuses on a particular region of operation

Global explanation considers the entire model

Algorithmic Understanding

A more technically inclined user or model builder may have different requirements

Properties of algorithm used

Whether all inputs to the algorithm seem useful and are understandable

User Explainable

A user should be able to generate explanations about how the algorithm works

Why Intelligibility matters?

Trust

- High accuracy by machine learning models
  - Imply model's ability to closely mimic data generating process
  - May possess the property of low interpretability by humans => **Intelligibility**
- Complex models do not explain their prediction well, which can act as a barrier to their adoption
- For mission critical applications
  - Saving life of patients in ICU
  - Taking smart decisions while flight landing or in airspace
  - Trajectory prediction while rocket launch
  - Self driving cars
- Interpretability helps in understanding the role of each feature contributing to the final outcome
- Complexity of models may hinder such causal effects

- Difficult to quantify
- More than just performance
- Driven X miles with Y crashes
- When negative events happen

Key Contributions

- Propose high performance GA^2Ms (Generalized Additive Models with pairwise interactions) for state-of-the-art accuracy and intelligibility
- Present two case studies that uncover interesting patterns
- Demonstrate scalability of method to large datasets
- Demonstrate intelligibility on dataset-level and individual patient-level

Background

- In mid 1990s
    - Cost-Effective HealthCare (CEHC)
- Goal: estimate probability of death (POD) due to pneumonia
- Choice of models: Neural nets v/s logistic regression
    - AUC score: NN - 0.86 vs LR - 0.77
- Neural nets more accurate but less intelligible, hence discarded
- Careful Consideration: NN are too risky for use on real patients
    - Finally used logistic regression, since weights for asthma could easily be adjusted

---

- In another study [2], rule-based model discovered some counterintuitive results
- E.g: HasAsthma(x) => LowerRisk(x)
    - Asthma patients or pregnant women are less prone to death by pneumonia
    - But it's easy to remove rules producing such generalisation and are hence editable

Generalized Additive Models (GAMs)

- Relationships between the individual predictors and the dependent variable follow smooth patterns that can be linear or nonlinear
- We can estimate these smooth relationships simultaneously and then predict g(E(Y))) by simply adding them up

$$g(E[y]) = \beta_0 + \sum f_j(x_j)$$

Link Function

Smooth
Non-parametric
functions

- When f_j is linear, g is called Generalized Linear Model (GLM)
- Model is intelligible since contribution of each term is clearly visible

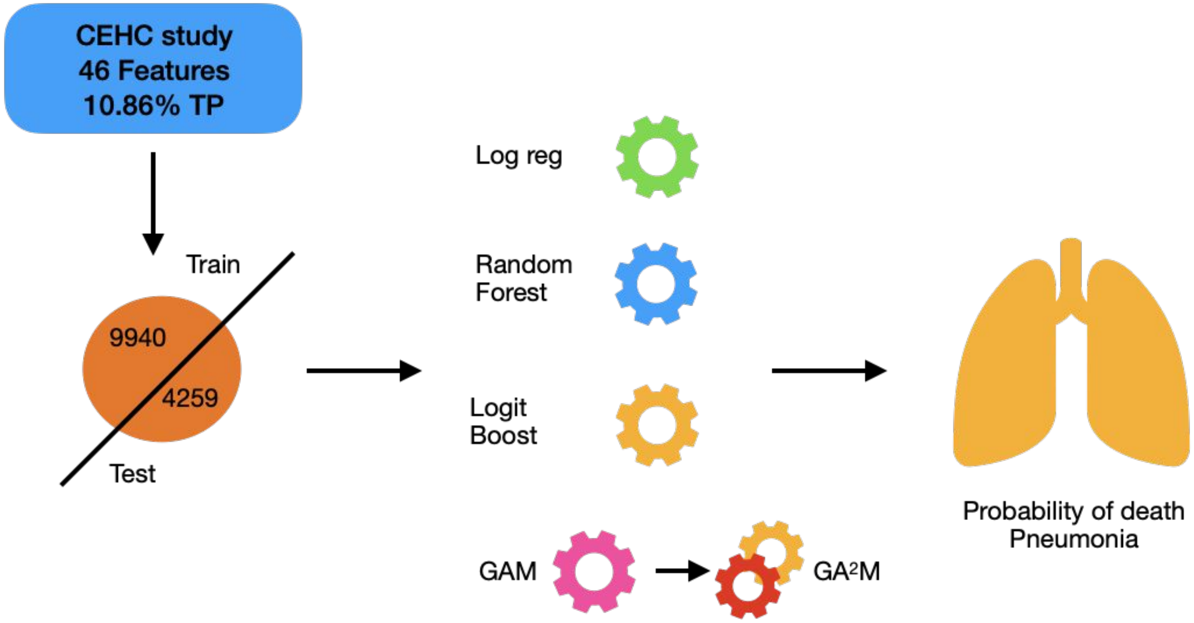
GAMs with pairwise interactions (GA^{2}Ms)

- Pairwise interactions added to improve accuracy

$$g(E[y]) = \beta_0 + \sum f_j(x_j) + \sum_{i \neq j} f_{ij}(x_i, x_j)$$

- Pairwise interactions can be represented using heat map and hence are intelligible
- GA^{2}M builds the best GAM and then detects and ranks all possible pairs of interactions in the residuals (includes top k pairs)
- Various methods to train GAMs and GA^{2}Ms - optimizing splines, regression
- Gradient boosting with bagging of shallow regression trees yields best accuracy

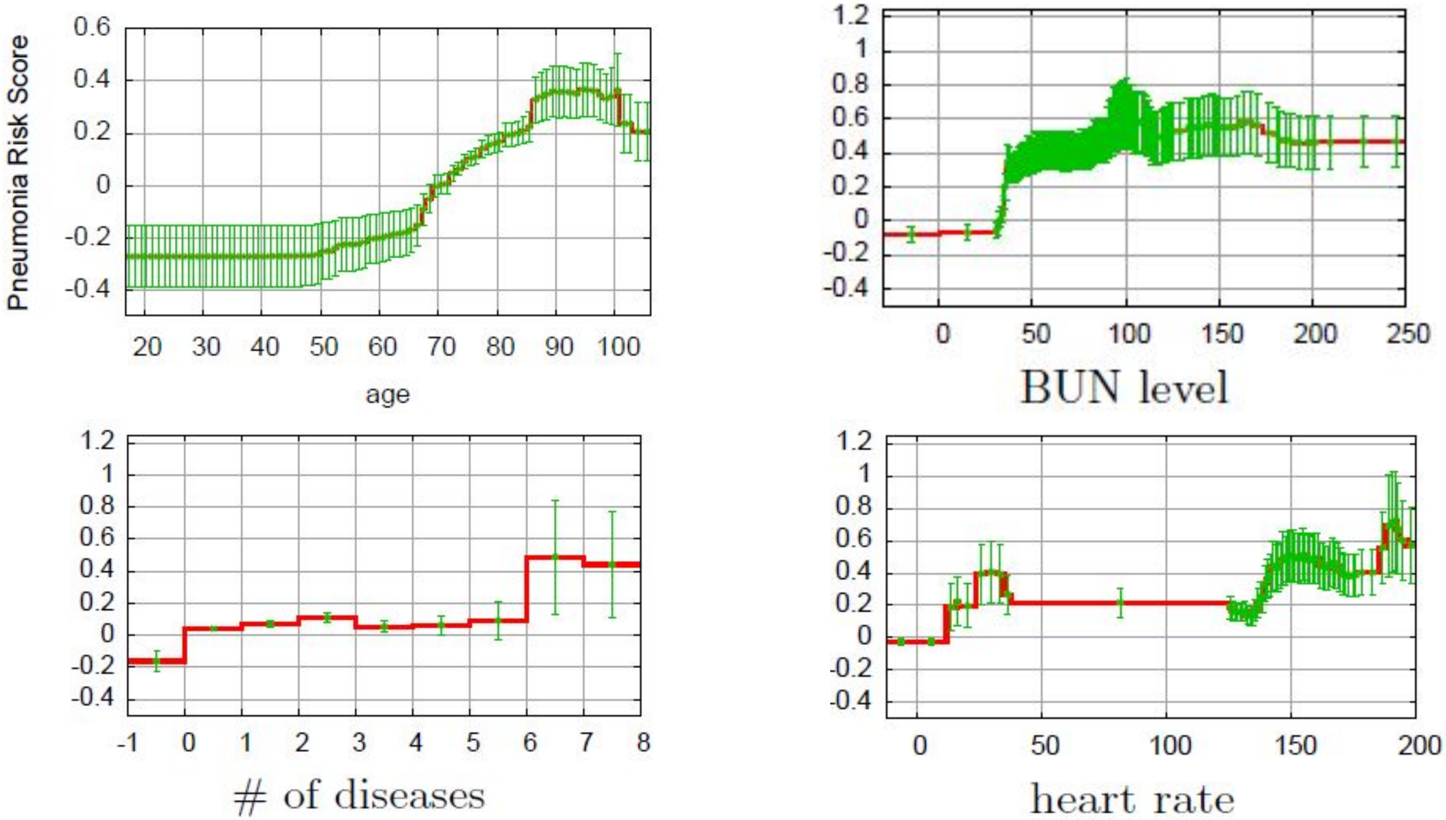
Case Study 1: Pneumonia Risk



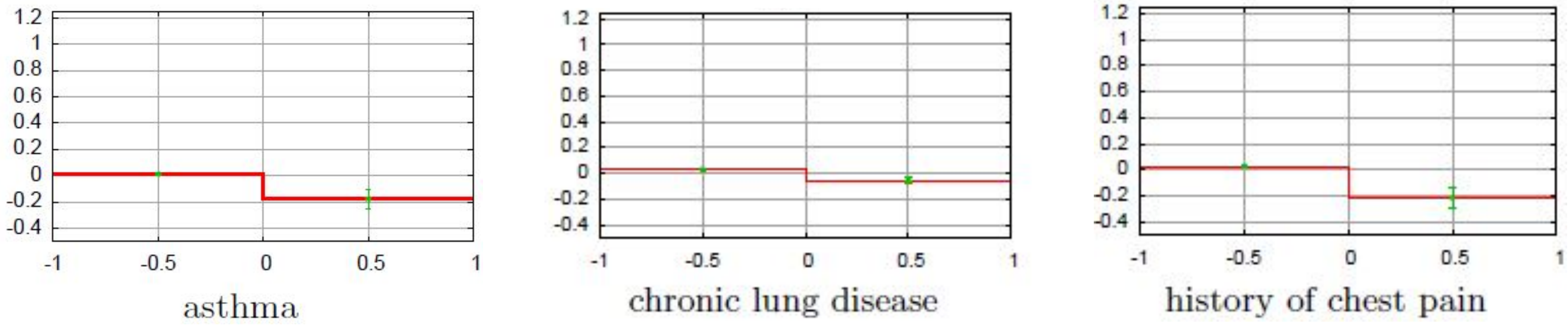
Model	Pneumonia
Logistic Regression	0.8432
GAM	0.8542
GA^2M	0.8576
Random Forests	0.8460
LogitBoost	0.8493

- Each term in the model returns a risk score (log odds) that is added to the aggregate predicted risk
- Terms with risk scores above zero increase risk; terms with scores below zero decrease risk

Observations



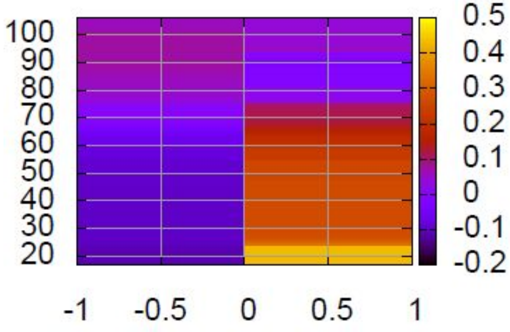
Observations



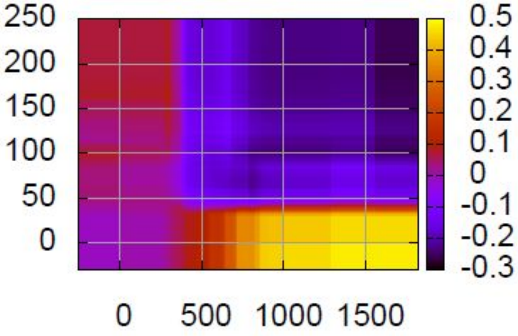
These disparities can be corrected by:

- Eliminating the terms from the model
- Using human expertise to redraw the graphs so that the risk score for condition=1 is positive, not negative

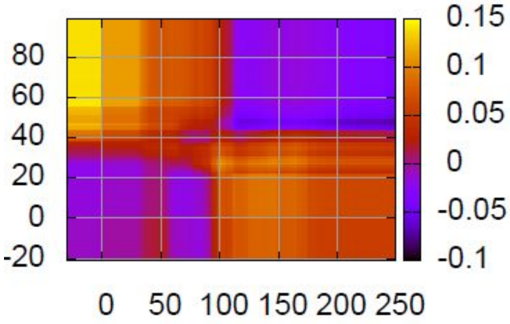
Pairwise Interactions



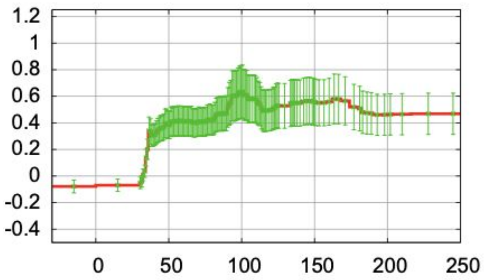
age vs. cancer



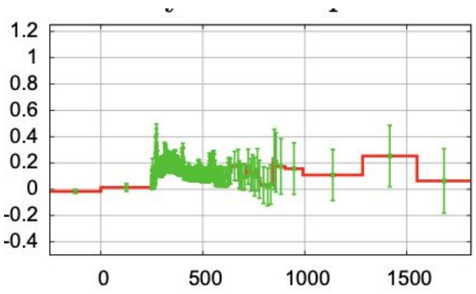
BUN vs. glucose



respiration rate vs. BUN

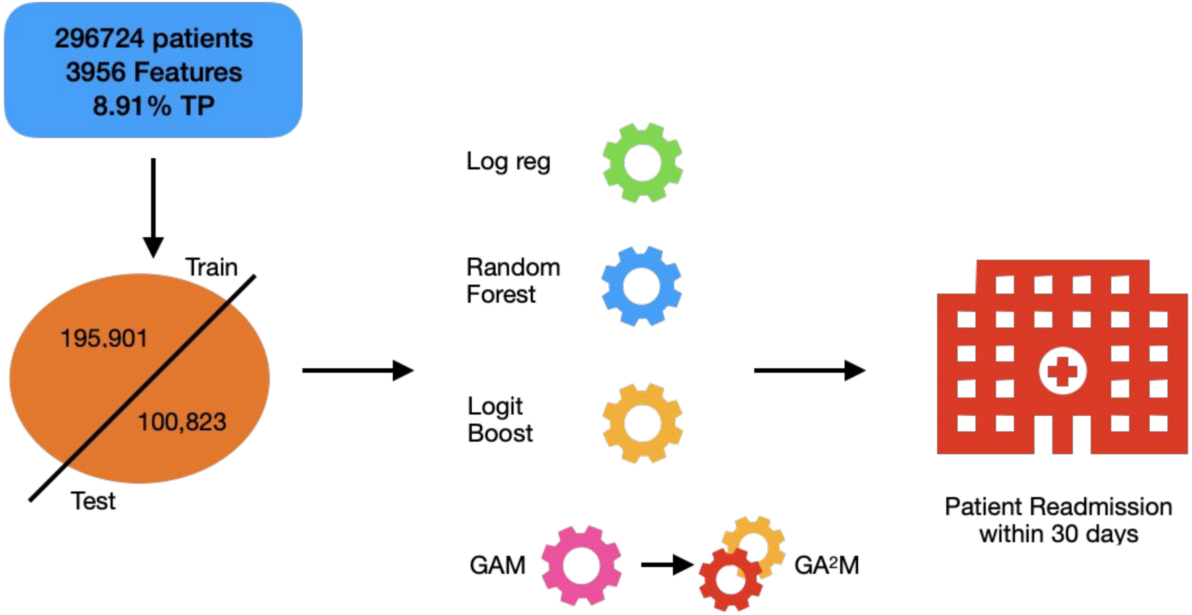


BUN level



glucose level

Case Study 2: 30-Day Readmission

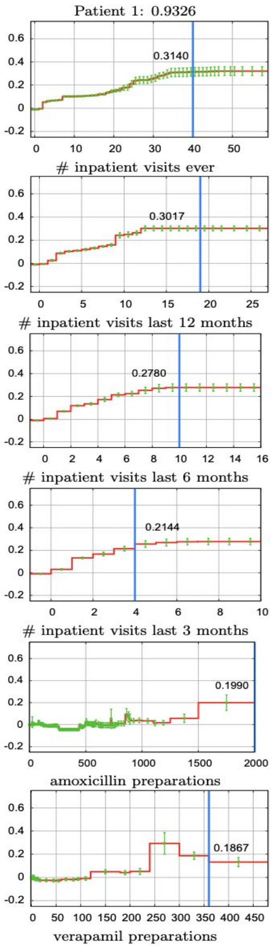


Model	Readmission
Logistic Regression	0.7523
GAM	0.7795
GA^2M	0.7833
Random Forests	0.7671
LogitBoost	0.7835

- Reasons for readmission - 1) Released the patient prematurely 2) Lack of adequate instructions 3) Lack of adequate follow-up
- Examine the predictions made by the model for three patients, instead of full model

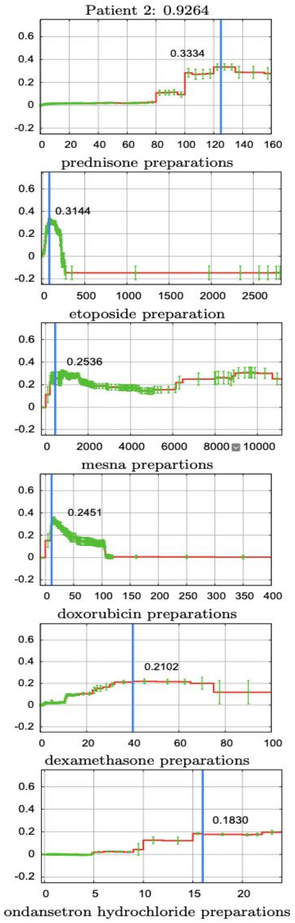
Observations (p = 0.9326)

- Features ranked according to the risk they contribute to that patient
- The terms that contribute most to their high probability of readmission are:
  - Total number of visits to the hospital
  - Large doses of
    - Amoxicillin (antibiotic used to treat infections like strep and pneumonia)
    - Verapamil (treatment for hypertension and angina), i.e., patient has an ongoing infection that may not be responding to antibiotics, and also probably has heart disease



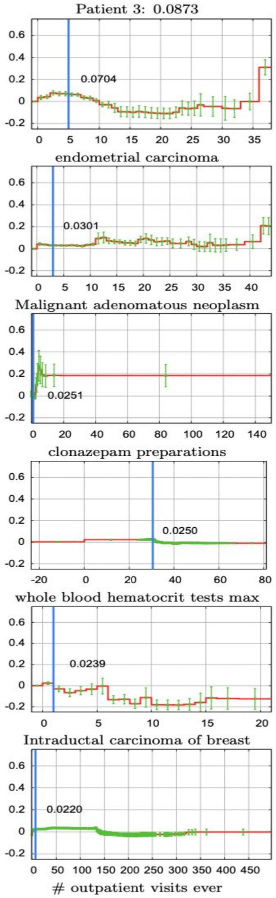
Observations (p = 0.9264)

- Features ranked according to the risk they contribute to that patient
  - prednisone - immuno suppressant
  - etoposide - anticancer drug
  - mesna - cancer chemotherapy drug
  - doxorubicin - treatment for blood and skin cancers
  - dexamethosone - immuno suppressant steroid
  - ondansetron - drug to treat nausea from chemotherapy
- Aggressive chemotherapy - High doses of these preparations suggest that cancer may not be responding well to treatment
- The contribution to risk from these 6 terms alone is greater than +1.5



Observations (p = 0.0873)

- Features ranked according to the risk they contribute to that patient
    - Endrometrial carcinoma - cancer common in post-menopausal women that can be treated by hysterectomy without radiation or chemotherapy
    - Benign abdominal tumor (val = 3)
    - Relaxant typically prescribed to calm patients or reduce spasms
    - Fairly typical (i.e. low risk) hematocrit test result
    - Pre-cancerous non-invasive lesion in the breast
    - Small number of outpatient visits (receiving treatment as outpatient without needing to be hospitalized)
- Patient has post-menopausal cancer that responds well to treatment if caught early, the treatments themselves are relatively low-risk, and didn't need unusual medications or hospitalization often in the last year



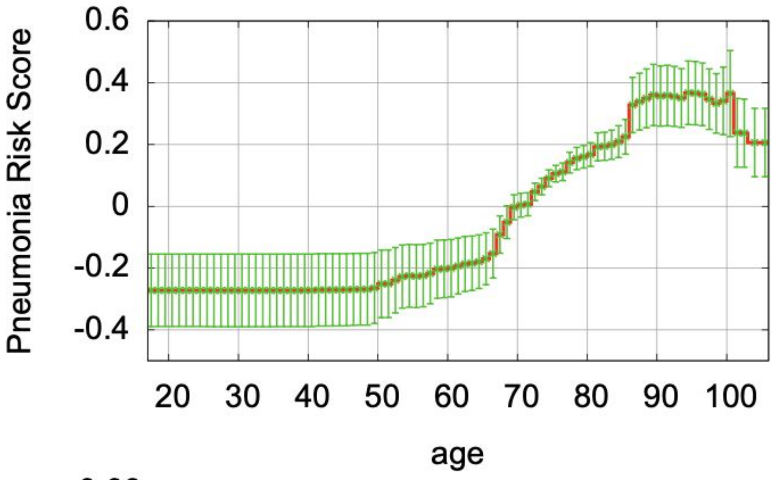
Discussion

- Sorting terms by importance
  - Ordering features quickly identifies the key patient characteristics that best explain the model's prediction
  - Help experts quickly understand the patient's condition
- Risk as a function of age
  - Present in both data sets and measured in years
  - In pneumonia: it explain why a patient has acquired pneumonia
  - In 30-day all-cause readmission: however, age is just one of thousands of factors

Age

Case Study 1

- Lower age significantly reduces the risk
  - 18-50: Risk is low and constant
  - 50-66: Rises slowly
  - 66-90: Quickly rises
  - 90 above: Levels off
- There is a small jump in risk
  - age 67
  - age 86
- Many patients would have retired at around age 65
- Differences in activity levels, health insurance, and willingness to get access healthcare early enough to improve outcomes
- Practitioners treat patients differently



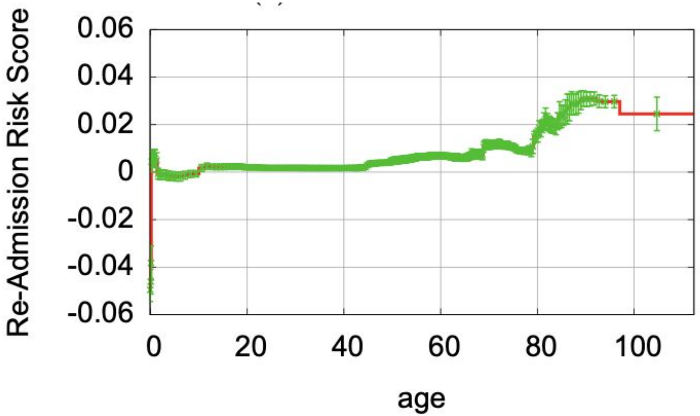
Age

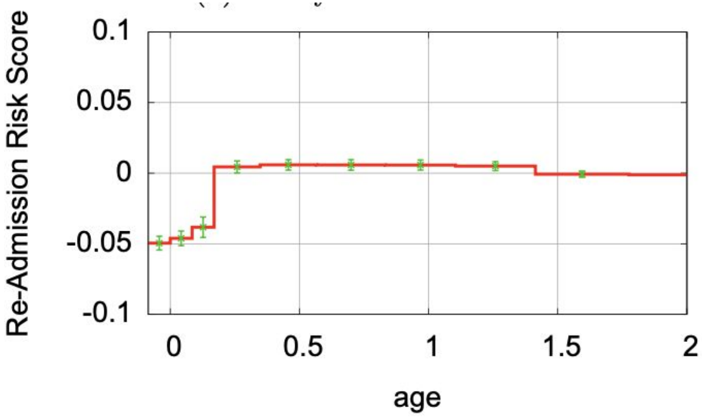
Case Study 2

- Dataset contains patients of all age
    - Including newborn infants

- Largest increase in score is +0.03 at age 90 and above

- x-axis has been expanded to show age 0-2 years
    - Newborns would not be discharged if they were at risk, the risk score for newborns aged 0-2 months is -0.04
    - Infants aged 3 - 15 months have higher risk





Key Takeaways

- Case studies demonstrate that the GA^2M models are intelligible
  - Macro level
  - Micro level
- Makes them suitable for deployment in the healthcare domain where applications demand debuggability and verification of results
- Easily scalable to large datasets

Observed Limitations

- Compete with ensemble techniques on dataset evaluated
  - Generalizability for explaining other complex tasks is questionable
- Propensity to overfit the data
- No prediction - Input is outside the trained data range
- Causality

Causality

Correlation does not imply causation

- It is tempting to interpret results causally
- What do we mean by Causality?
  - Patient has X => Received treatments A, B, and C and
  - Noting amount of A, B, and C patient received => Patient is not responding well
- Instead, GA^2M learns
  - high a doses of A, B, and C are associated with high risk or readmission
- Upto experts to infer the underlying causal reasons for the feature values and the risk they predict

Questions?

References:

[1] URL “https://www.borealisai.com/en/blog/intelligibility-key-component-trust-machine-learning/”

[2] R. Ambrosino, B. Buchanan, G. Cooper, and M. Fine. The use of misclassification costs to learn rule-based decision support models for cost-effective hospital admission strategies. In Proceedings of the Annual Symp. on Comp. Application in Medical Care, 1995.