UNIVERSITY of York

ASSURING AUTONOMY INTERNATIONAL PROGRAMME

The impact of training data shortfalls on diabetes comorbidity predictor

Dr. Philippa Ryan, Mr. Berk Ozturk, Dr. Tom Lawton, Prof. Ibrahim Habli

Introduction

- Development of diabetes comorbidity predictor
  - Provides ”independent second opinion” on patient
  - Example for hypertension
- Training issues and safety analysis of DCP
- Results and next steps

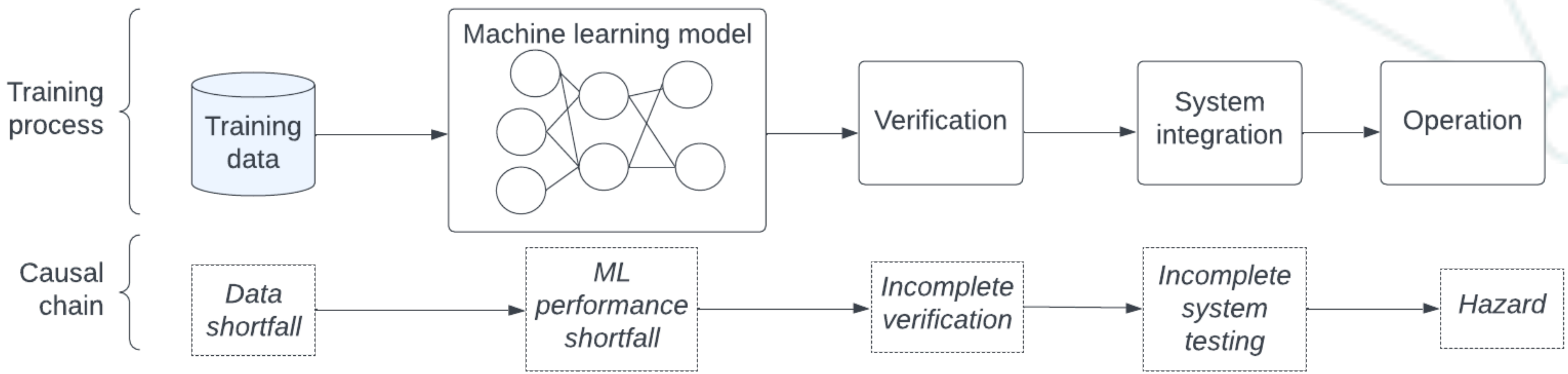
- Funded by
  - EPSRC Assuring Responsibility-Trustworthy Autonomous Systems
  - LRF Assuring Autonomy International Programme

Ryan P, Ozturk, B., Lawton, T., Habli, I.: The Impact of Training Data Shortfalls on Safety of AI-based Clinical Decision Support Systems. *In: SAFECOMP 2023 (to appear Sept.)*

Training Data for ML

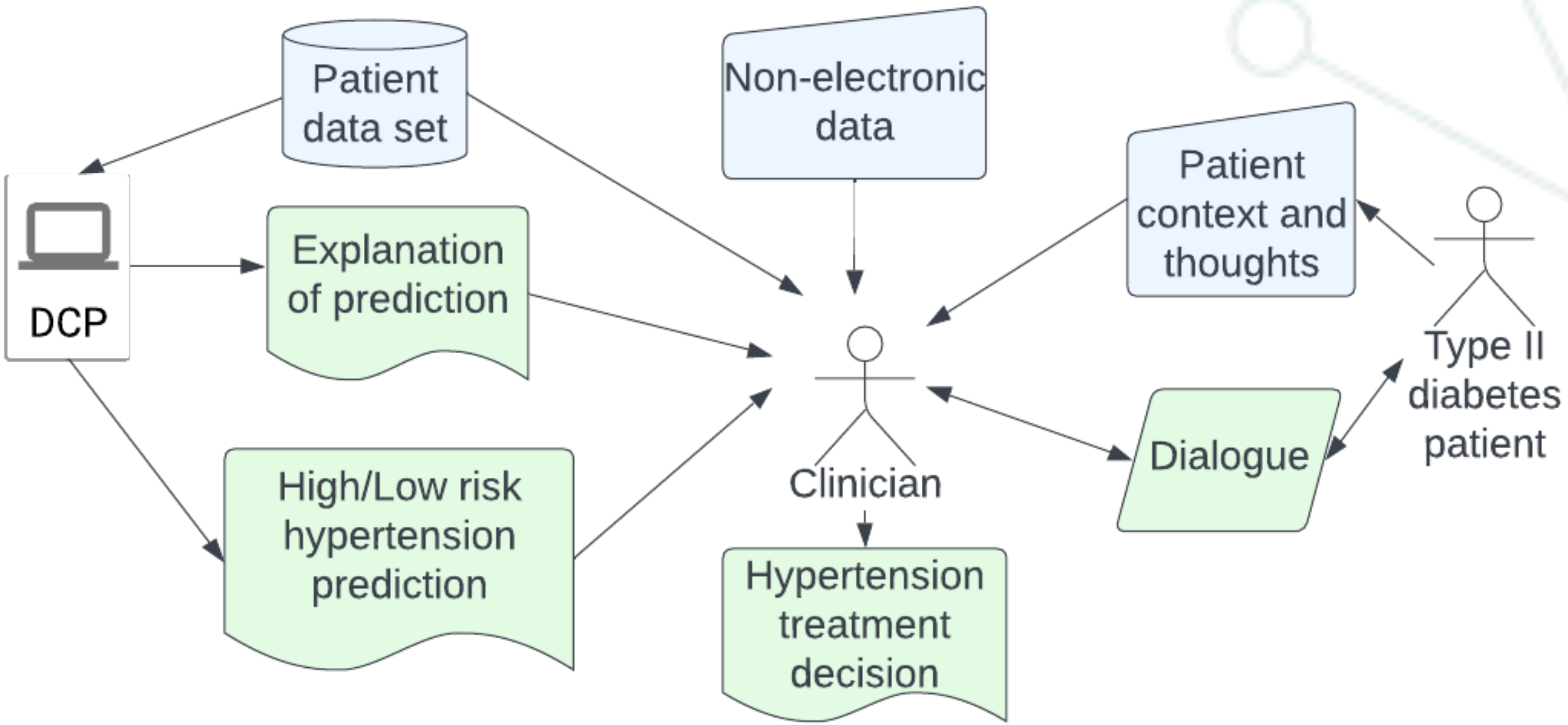
- All Machine Learning needs good quality training data
  - Data embodies the functionality you want it to learn
  - User generated data
    - Issues with validity (values, representative of reality)
    - Better for coverage (generate cases)
  - Real world datasets
    - Fewer issues with validity for individual data points
    - Harder to argue future coverage and distribution
- Any problems with training data reflected in final ML

Latent failures



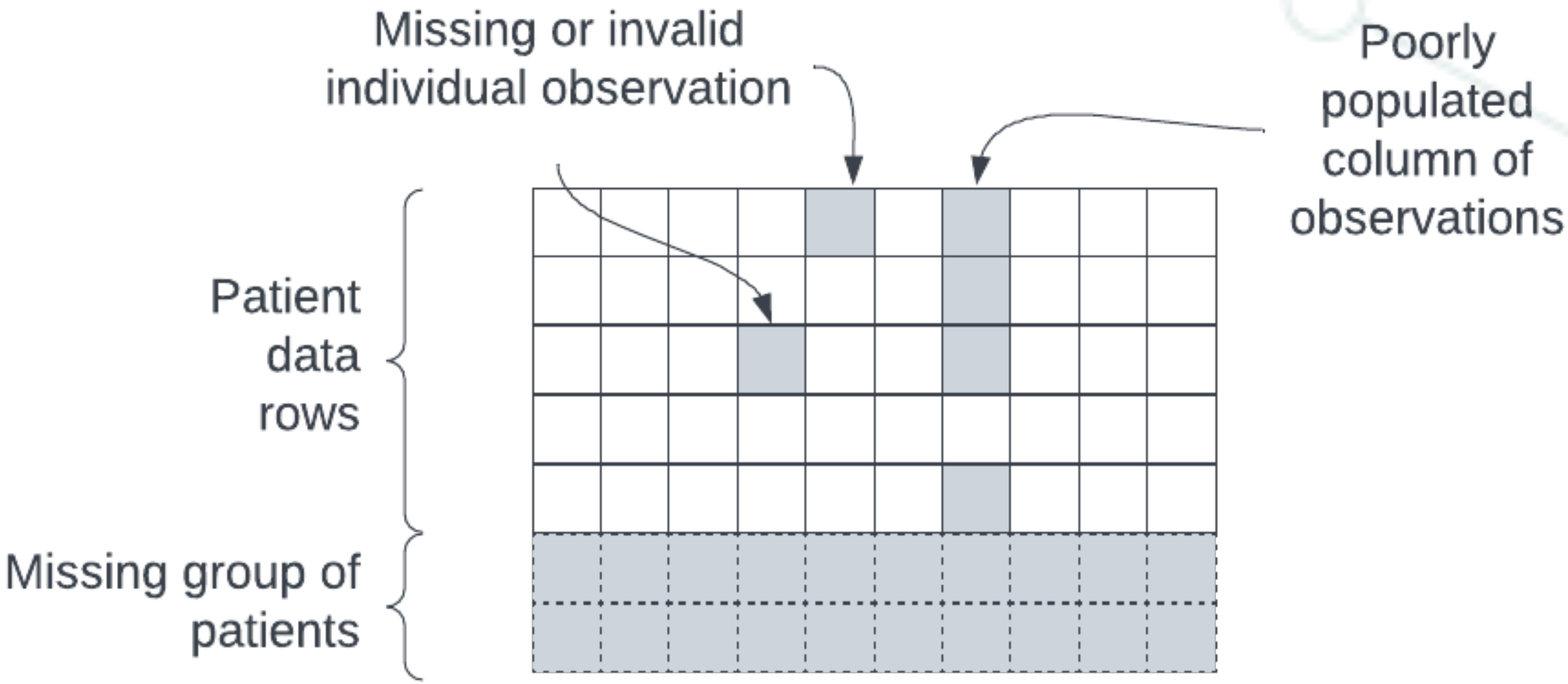
DCP use case

Hypertension version



The training data

Connecting Bradford - database



- 43,000+ data training rows used of Type II Diabetes patients
- Reduced feature space (14,000+) to 20 FOI
    - Reviewed by clinician for validation

What can we do?

Can pre-process real-world data

- Missing values common problem with medical diagnosis ML
- Can compensate => data imputation
  - Lots of methods e.g., average, median
  - Bag imputation
    - Uses ML to predict likely values for missing cases
- But can introduce bias
- A lots of research maximises metrics without understanding risk

Hazards

Hypertension version

- DCP output could influence decision
- False positive
  - Patient categorised high risk when they are not
  - Provided with medication they don’t need with side-effects (severe)
- False negative
  - Patient categorised low risk when they are not
  - Risk of heart attack/stroke (catastrophic)
- Likelihood of incorrect diagnosis from DCP hard to predict
  - Varies per patient

Safety analysis

Hazop like

- “Flow” – training data into the training process
- Guideword examples:
  - More - indicates a bias in the data, e.g., over representation of particular patient group in the dataset
  - No or Not - FOI or set of FOIs are missing
  - Less - fewer examples of FOI than are desirable for good performance are present
  - Early/Before - indicates that a FOI may be present but out of date with respect to the co-morbidity presenting itself
  - Reverse – opposite diagnosis included

Guideword	Deviation	Cause	Effect	Mitigation
No or not	Samples for ethnic group not included in training data (TD)	No/limited patients of ethnic group were patients	ML not trained or verified adequately for ethnic group with higher genetic risk of hypertension	Manual review of DB by expert, show clinician prototypical examples, patient discussion
Part of	Partially missing BMI in TD samples	BMI not consistently recorded	ML performance biased based on the data imputation method used, leads to poor performance for high or low BMI patients	Use bag imputation for TD records to reduce bias, recommend collection of BMI for future TD, show clinician prototype examples, patient discussion
More	Over representation in TD of high BMI patients	Most patients examined had high BMI	Prediction biased towards patients with high BMI, meaning patients with low BMI have less accurate predictions	Manual review of DB by expert, training samples picked across all ranges, show clinician prototype examples, patient discussion
More	Over representation in TD of certain ethnic group	Over diagnosis by trained ML for patients of other ethnic groups	TD dominated by ethnic group with genetic disposition to hyper tension	Manual review of DB by expert, show clinician prototype examples, patient discussion
Early/ Before and More	BMI data is out of date and training patients have changed BMI by time of diagnosis	DB not kept up to date, TD sampled from wrong part of patient history	ML underestimates likelihood of hypertension	TD selected from samples near to hypertension diagnosis, manual review of DB by expert, patient discussion
Instead	BMI value no longer highest FOI for some FOI distribution	Performance outlier from ML	Wrong prediction for hypertension	Show clinician FOI from training and for each prediction at point of use, patient discussion

Discussion

- Prototypical examples
  - Issue of patient confidentiality
  - Would need to obfuscate these further
- Limited to 20 FOI during training may miss data patterns
  - Some FOI result of hypertension not cause
- Missing data can be significant
  - Patient too unwell for tests
  - Long term trend in their health
  - Or could just be poor record keeping!
  - How do we incorporate in ML process?
- Scalability
  - How to perform manual review of such a large set of data?

Summary

- Issues with training data lead to latent ML faults
    - Subtle and varied
- Too many papers in this area maximise metrics without understanding risk
- System level hazard analysis
    - Can help identify actual risk with more clarity
    - We can put *targeted* mitigations in place
- May be complex trade-offs

UNIVERSITY of York

ASSURING AUTONOMY

INTERNATIONAL PROGRAMME

Funded by

Lloyd's Register Foundation

UNIVERSITY of York

Training DCP

- Data selection
  - 42,000+ data training rows used of Type II Diabetes patients
  - Reduced feature space (14,000+) to 20 FOI
    - Reviewed by clinician for validation
  - Removed duplicate records
  - Normalised values
  - Compensated missing values using bag imputation
- Trained multiple ML models
  - Naïve Bayes, NN, random forest, SVM
- Ensemble gave best results
  - Accuracy and Kappa values
- NICE guidelines used

Ozturk, B., Lawton, T., Smith, S., Habli, I.: Predicting Progression of Type 2 Diabetes using Primary Care Data with the Help of Machine Learning. In: Medical Informatics Europe 2023 (2023)

Feature Importance Levels