Machine learning approaches to (i) predicting response to therapy in diabetes, (ii) data-driven diabetes subtype classification and (iii) synthetic data generation

Chris Sainsbury
Queen Elizabeth University Hospital, Glasgow
NRS Senior Fellow
Hon Senior Research Fellow, University of Birmingham

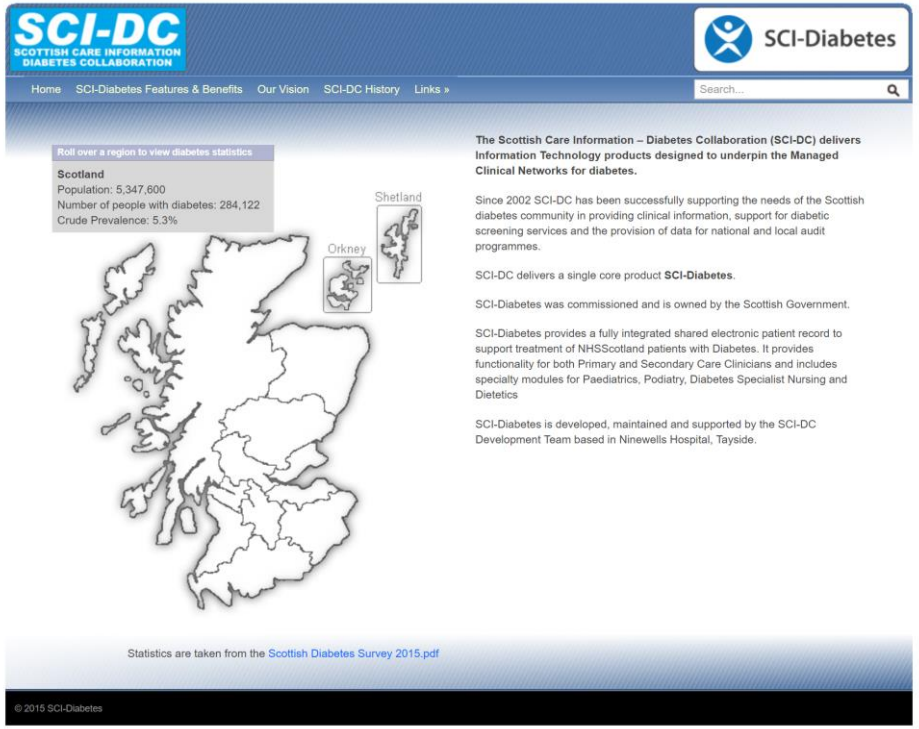
@csainsbury
http://glucose.ai
http://github.com/csainsbury/

NHS Greater Glasgow and Clyde
NHS SCOTLAND NHS RESEARCH SCOTLAND
UNIVERSITY OF BIRMINGHAM
kvatchii.com intelligent health
Innovate UK

- MD: endothelial function
- mathematical modelling of systemic/pulmonary circulation
- inpatient CBG data analysis
    - patient risk stratification
    - system performance analysis (clinical unit / individual operator)
    - just-in-time education for ward staff
- SCI diabetes data analysis
    - HbA1c variability and outcome in T1DM
    - multiscale / multiparameter variability and outcome
- Machine Learning
    - ANN vs logistic regression vs clinician prediction of DM type at diagnosis
    - myDiabetesIQ
    - THINKINGAI group – University of Birmingham

SCI-Diabetes

‘we have (one of) the best datasets in the world, so there is an opportunity to build high quality algorithms to understand relationships within diabetes data…’



what questions to address?

initial questions / prediction problems chosen

i. suggest best next therapy / combination of therapies to achieve goals in multiple domains (hba1c reduction / blood pressure / mortality etc)

ii. predict complications (LLA, CV events)

iii. predict acute complications (hypoglycaemia etc)

iv. predict diabetes type at diagnosis, identify MODY etc

Innovate UK (Digital Health Technology Catalyst)
1M grant 2018-2021

Funding competition

Digital health technology catalyst 2017 round 1

UK businesses can apply for a share of up to £8 million to speed up development of new digital technology healthcare solutions.

Competition opens: Monday 31 July 2017

Competition closes: Wednesday 11 October 2017 12:00pm

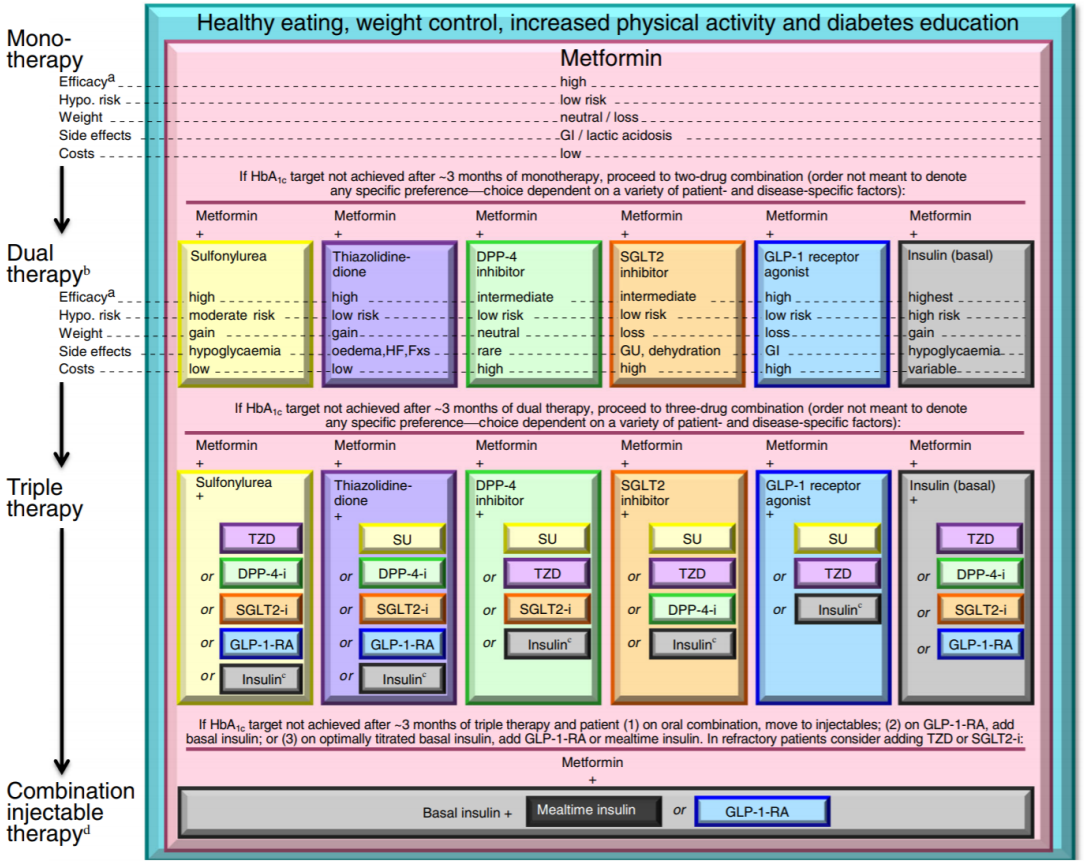
Competition: Digital Health Technology Catalyst 2017 Round 1
Project Title: MyDiabetesIQ

Question 1

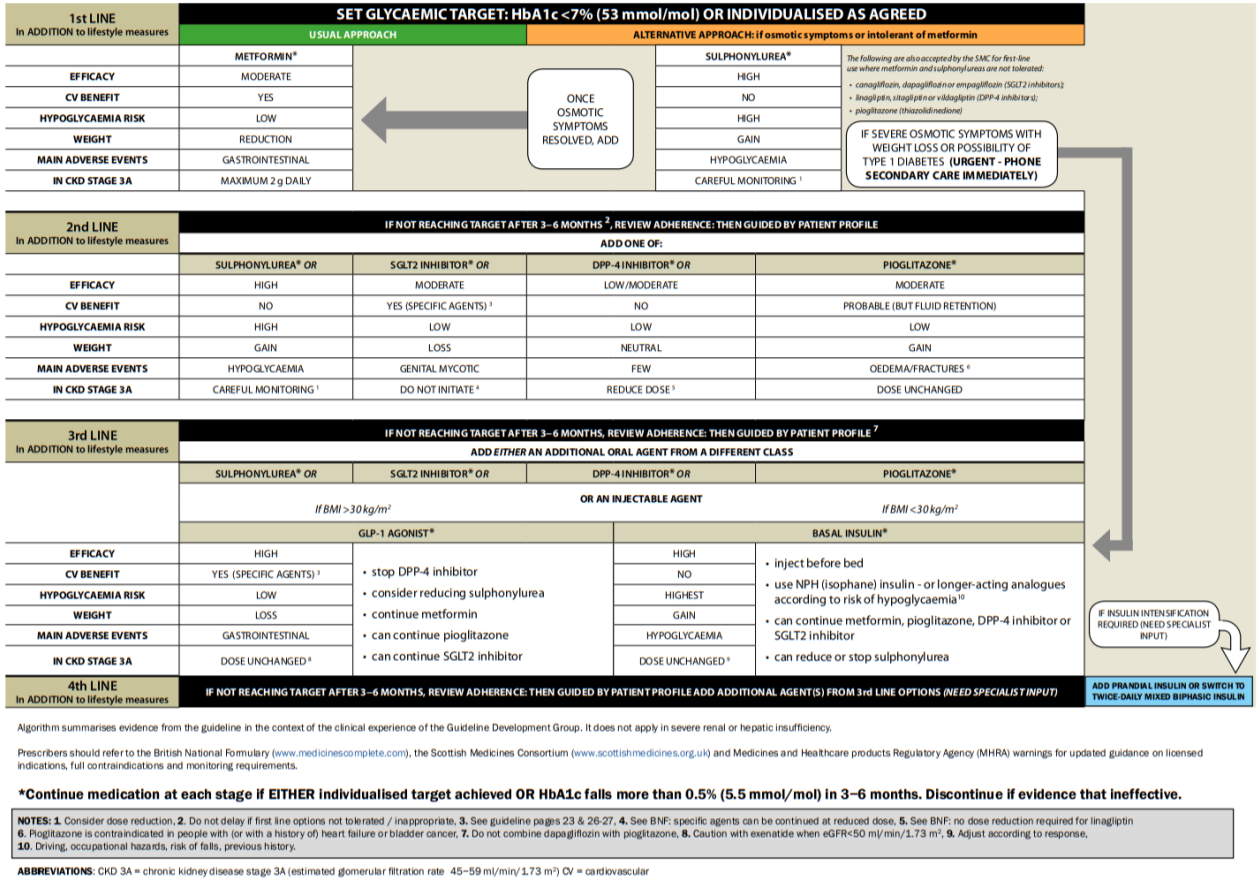
To generate a prediction of an individual’s response to any given clinically appropriate drug / combination of drugs - drawing on insight gained from the population over time.

or

what is/are the next best drug(s) for my patient?



ADA / EASD, 2015



1st LINE In ADDITION to lifestyle measures	SET GLYCAEMIC TARGET: HbA1c <7% (53 mmol/mol) OR INDIVIDUALISED AS AGREED		
	USUAL APPROACH	ALTERNATIVE APPROACH: if osmotic symptoms or intolerant of metformin	
	METFORMIN*	SULPHONYLUREA*	The following are also accepted by the SMC for first-line use where metformin and sulphonylureas are not tolerated: • canagliflozin, dapagliflozin or empagliflozin (SGLT2 inhibitors); • linagliptin, sitagliptin or vildagliptin (DPP-4 inhibitors); • pioglitazone (thiazolidinedione)
EFFICACY	MODERATE	HIGH	
CV BENEFIT	YES	NO	
HYPOGLYCAEMIA RISK	LOW	HIGH	
WEIGHT	REDUCTION	GAIN	
MAIN ADVERSE EVENTS	GASTROINTESTINAL	HYPOGLYCAEMIA	
IN CKD STAGE 3A	MAXIMUM 2 g DAILY	CAREFUL MONITORING [1]	

ONCE OSMOTIC SYMPTOMS RESOLVED, ADD

IF SEVERE OSMOTIC SYMPTOMS WITH WEIGHT LOSS OR POSSIBILITY OF TYPE 1 DIABETES **(URGENT - PHONE SECONDARY CARE IMMEDIATELY)**

2nd LINE In ADDITION to lifestyle measures	IF NOT REACHING TARGET AFTER 3–6 MONTHS [2], REVIEW ADHERENCE: THEN GUIDED BY PATIENT PROFILE			
	ADD ONE OF:			
	SULPHONYLUREA* *OR*	SGLT2 INHIBITOR* *OR*	DPP-4 INHIBITOR* *OR*	PIOGLITAZONE*
EFFICACY	HIGH	MODERATE	LOW/MODERATE	MODERATE
CV BENEFIT	NO	YES (SPECIFIC AGENTS) [3]	NO	PROBABLE (BUT FLUID RETENTION)
HYPOGLYCAEMIA RISK	HIGH	LOW	LOW	LOW
WEIGHT	GAIN	LOSS	NEUTRAL	GAIN
MAIN ADVERSE EVENTS	HYPOGLYCAEMIA	GENITAL MYCOTIC	FEW	OEDEMA/FRACTURES [6]
IN CKD STAGE 3A	CAREFUL MONITORING [1]	DO NOT INITIATE [4]	REDUCE DOSE [5]	DOSE UNCHANGED

3rd LINE In ADDITION to lifestyle measures	IF NOT REACHING TARGET AFTER 3–6 MONTHS, REVIEW ADHERENCE: THEN GUIDED BY PATIENT PROFILE [7]				
	ADD *EITHER* AN ADDITIONAL ORAL AGENT FROM A DIFFERENT CLASS				
	SULPHONYLUREA* *OR*	SGLT2 INHIBITOR* *OR*	DPP-4 INHIBITOR* *OR*	PIOGLITAZONE*	
		If BMI >30 kg/m²	OR AN INJECTABLE AGENT	*If BMI <30 kg/m²*	
	GLP-1 AGONIST*		BASAL INSULIN*		
EFFICACY	HIGH	• stop DPP-4 inhibitor • consider reducing sulphonylurea • continue metformin • can continue pioglitazone • can continue SGLT2 inhibitor	HIGH	• inject before bed • use NPH (isophane) insulin - or longer-acting analogues according to risk of hypoglycaemia [10] • can continue metformin, pioglitazone, DPP-4 inhibitor or SGLT2 inhibitor • can reduce or stop sulphonylurea	
CV BENEFIT	YES (SPECIFIC AGENTS) [3]		NO		
HYPOGLYCAEMIA RISK	LOW		HIGHEST		
WEIGHT	LOSS		GAIN		
MAIN ADVERSE EVENTS	GASTROINTESTINAL		HYPOGLYCAEMIA		
IN CKD STAGE 3A	DOSE UNCHANGED [8]		DOSE UNCHANGED [9]		

IF INSULIN INTENSIFICATION REQUIRED (NEED SPECIALIST INPUT)

4th LINE In ADDITION to lifestyle measures	IF NOT REACHING TARGET AFTER 3–6 MONTHS, REVIEW ADHERENCE: THEN GUIDED BY PATIENT PROFILE ADD ADDITIONAL AGENT(S) FROM 3rd LINE OPTIONS *(NEED SPECIALIST INPUT)*	ADD PRANDIAL INSULIN OR SWITCH TO TWICE-DAILY MIXED BIPHASIC INSULIN

Algorithm summarises evidence from the guideline in the context of the clinical experience of the Guideline Development Group. It does not apply in severe renal or hepatic insufficiency.

Prescribers should refer to the British National Formulary (www.medicinescomplete.com), the Scottish Medicines Consortium (www.scottishmedicines.org.uk) and Medicines and Healthcare products Regulatory Agency (MHRA) warnings for updated guidance on licensed indications, full contraindications and monitoring requirements.

***Continue medication at each stage if EITHER individualised target achieved OR HbA1c falls more than 0.5% (5.5 mmol/mol) in 3–6 months. Discontinue if evidence that ineffective.**

NOTES: 1. Consider dose reduction, **2.** Do not delay if first line options not tolerated / inappropriate, **3.** See guideline pages 23 & 26-27, **4.** See BNF: specific agents can be continued at reduced dose, **5.** See BNF: no dose reduction required for linagliptin **6.** Pioglitazone is contraindicated in people with (or with a history of) heart failure or bladder cancer, **7.** Do not combine dapagliflozin with pioglitazone, **8.** Caution with exenatide when eGFR<50 ml/min/1.73 m², **9.** Adjust according to response, **10.** Driving, occupational hazards, risk of falls, previous history.

ABBREVIATIONS: CKD 3A = chronic kidney disease stage 3A (estimated glomerular filtration rate 45–59 ml/min/1.73 m²) CV = cardiovascular

SIGN 154, 2018

what is the next best drug(s) for my patient?
virtual n = 1 drug trial

eg what drug should I prescribe to give this patient the best chance of having an HbA1c <60mmol/mol, with a reduction in blood pressure and BMI in 1 year?

taking into account their individual history of:

- HbA1c / BMI / blood pressure
- previously prescribed combinations of drug therapies
- how previous drugs have impacted on HbA1c / BMI / blood pressure

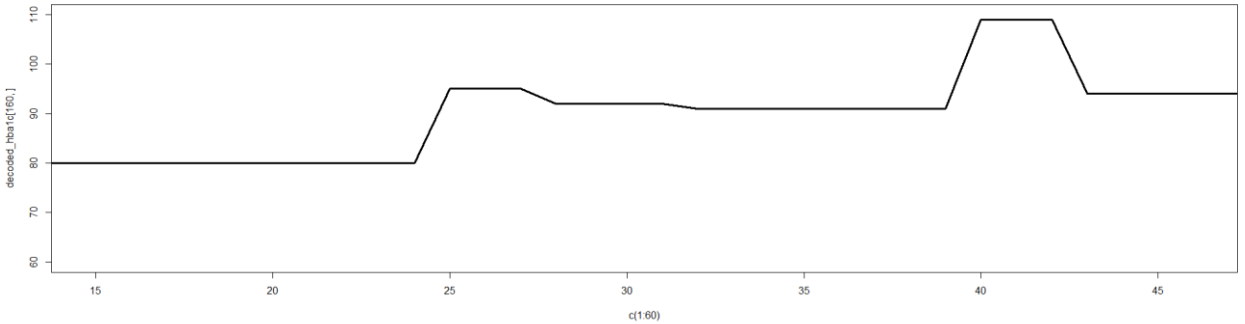
time series

- sex
- age
- ethnicity

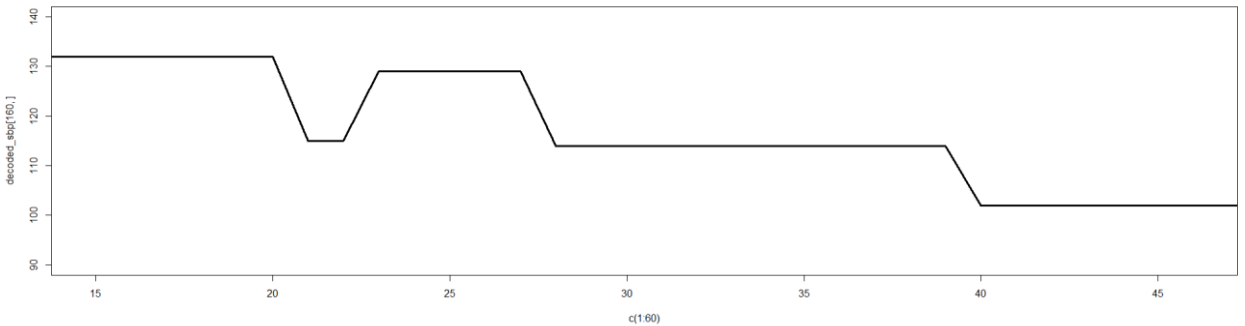
stable over time

managing time series data – numerical data

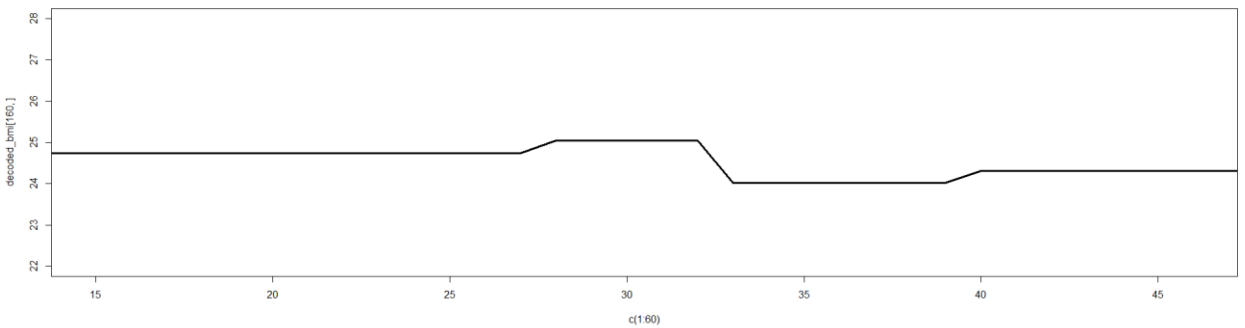
hba1c



sbp

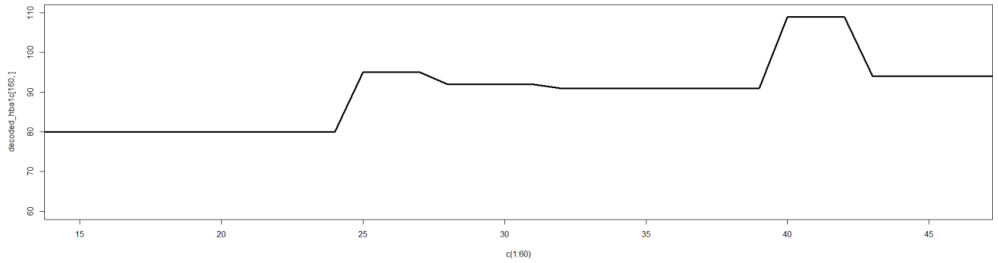


bmi



time

managing time series data – 2
numerical time series



managing missingness

- imputation
  - lvcf / other
  - interpolation – spline / linear
- masking

scaling / normalisation

RECURRENT NEURAL NETWORKS FOR MULTIVARIATE TIME SERIES WITH MISSING VALUES

Zhengping Che, Sanjay Purushotham
Department of Computer Science
University of Southern California
Los Angeles, CA 90089, USA
{zche,spurusho}@usc.edu

Kyunghyun Cho, David Sontag
Department of Computer Science
New York University
New York, NY 10012, USA
kyunghyun.cho@nyu.edu,dsontag@cs.nyu.edu

Yan Liu
Department of Computer Science
University of Southern California
Los Angeles, CA 90089, USA
yanliu.cs@usc.edu

ABSTRACT

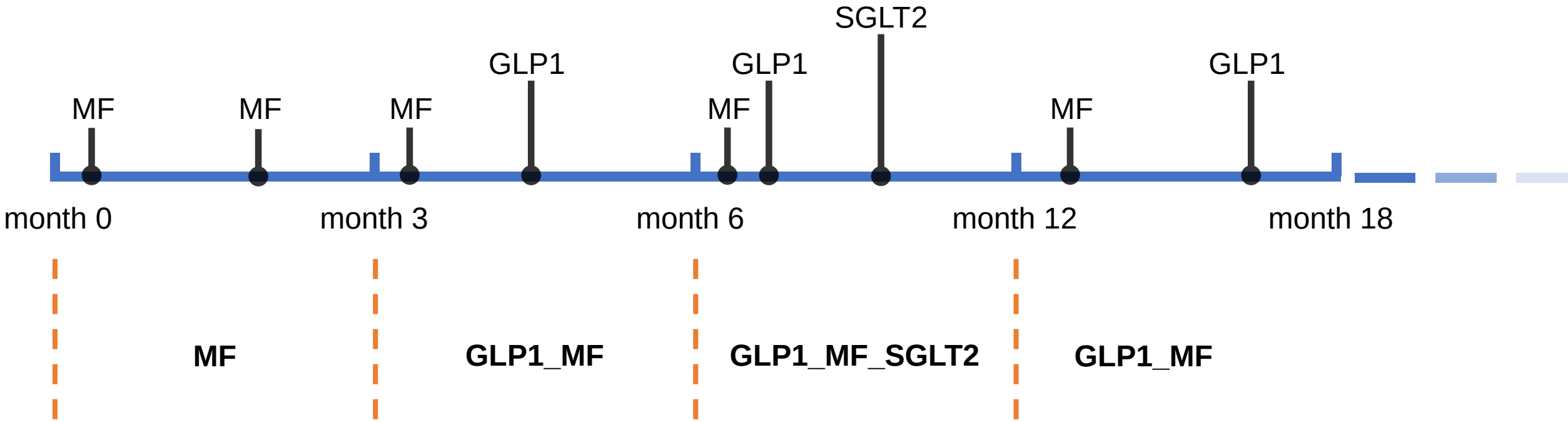
Multivariate time series data in practical applications, such as health care, geoscience, and biology, are characterized by a variety of missing values. In time series prediction and other related tasks, it has been noted that missing values and their missing patterns are often correlated with the target labels, a.k.a., *informative* missingness. There is very limited work on exploiting the missing patterns for effective imputation and improving prediction performance. In this paper, we develop novel deep learning models, namely GRU-D, as one of the early attempts. GRU-D is based on Gated Recurrent Unit (GRU), a state-of-the-art recurrent neural network. It takes two representations of missing patterns, i.e., *masking* and *time interval*, and effectively incorporates them into a deep model architecture so that it not only captures the long-term temporal dependencies in time series, but also utilizes the missing patterns to achieve better prediction results. Experiments of time series classification tasks on real-world clinical datasets (MIMIC-III, PhysioNet) and synthetic datasets demonstrate that our models achieve state-of-the-art performance and provides useful insights for better understanding and utilization of missing values in time series analysis.

arXiv:1606.01865v2 [cs.LG] 7 Nov 2016

managing time series data – prescription data

```
2017-06-30,Lantus 100units/ml solution for injection 3ml pre-filled SoloStar pen (Sanofi),6.1.1.2
2017-09-15,NovoRapid FlexPen 100units/ml solution for injection 3ml pre-filled pen (Novo Nordisk Ltd),6.1.1.1
2017-05-08,Levemir FlexPen 100units/ml solution for injection 3ml pre-filled pen (Novo Nordisk Ltd),6.1.1.2
2017-06-30,Humalog KwikPen 100units/ml solution for injection 3ml pre-filled pen (Eli Lilly and Company Ltd),6.1.1.1
2016-12-05,Gliclazide 80mg tablets,6.1.2.1
2017-09-15,Levemir FlexPen 100units/ml solution for injection 3ml pre-filled pen (Novo Nordisk Ltd),6.1.1.2
2017-05-08,NovoRapid FlexPen 100units/ml solution for injection 3ml pre-filled pen (Novo Nordisk Ltd),6.1.1.1
2017-04-04,NovoRapid FlexPen 100units/ml solution for injection 3ml pre-filled pen (Novo Nordisk Ltd),6.1.1.1
2017-04-04,Levemir FlexPen 100units/ml solution for injection 3ml pre-filled pen (Novo Nordisk Ltd),6.1.1.2
2017-08-01,NovoRapid FlexPen 100units/ml solution for injection 3ml pre-filled pen (Novo Nordisk Ltd),6.1.1.1
2017-06-30,Gliclazide 80mg tablets,6.1.2.1
2017-03-06,NovoRapid FlexPen 100units/ml solution for injection 3ml pre-filled pen (Novo Nordisk Ltd),6.1.1.1
2017-03-06,Levemir FlexPen 100units/ml solution for injection 3ml pre-filled pen (Novo Nordisk Ltd),6.1.1.2
2017-06-30,Metformin 850mg tablets,6.1.2.2
2016-12-15,GlucoGel 40% gel original (BBI Healthcare Ltd),6.1.4.0
2017-01-30,NovoRapid Penfill 100units/ml solution for injection 3ml cartridges (Novo Nordisk Ltd),6.1.1.1
2016-11-04,NovoRapid Penfill 100units/ml solution for injection 3ml cartridges (Novo Nordisk Ltd),6.1.1.1
2017-01-30,Lantus 100units/ml solution for injection 3ml pre-filled SoloStar pen (Sanofi),6.1.1.2
2016-11-22,Metformin 850mg tablets,6.1.2.2
2017-04-24,Metformin 850mg tablets,6.1.2.2
2017-04-24,Gliclazide 80mg tablets,6.1.2.1
2016-11-04,Lantus 100units/ml solution for injection 3ml pre-filled SoloStar pen (Sanofi),6.1.1.2
2016-11-22,Gliclazide 80mg tablets,6.1.2.1
2016-12-16,GlucaGen Hypokit 1mg powder and solvent for solution for injection (Novo Nordisk Ltd),6.1.4.0
2017-03-07,Gliclazide 80mg tablets,6.1.2.1
2017-05-31,Lantus 100units/ml solution for injection 3ml pre-filled SoloStar pen (Sanofi),6.1.1.2
2017-01-06,Gliclazide 80mg tablets,6.1.2.1
2017-04-18,NovoRapid Penfill 100units/ml solution for injection 3ml cartridges (Novo Nordisk Ltd),6.1.1.1
2017-06-30,Gliclazide 80mg tablets,6.1.2.1
2017-03-07,Gliclazide 80mg tablets,6.1.2.1
2017-03-06,NovoRapid FlexPen 100units/ml solution for injection 3ml ...,6.1.1.1
2017-03-06,Levemir FlexPen 100units/ml solution for injection 3ml pr...,6.1.1.2
2016-11-04,NovoRapid Penfill 100units/ml solution for injection 3ml ...,6.1.1.1
2017-01-25,Lantus 100units/ml solution for injection 3ml pre-filled ...,6.1.1.2
2017-08-17,Gliclazide 80mg tablets,6.1.2.1
2017-05-24,Gliclazide 80mg tablets,6.1.2.1
2017-04-25,Gliclazide 80mg tablets,6.1.2.1
2017-03-28,Gliclazide 80mg tablets,6.1.2.1
2017-02-27,Gliclazide 80mg tablets,6.1.2.1
2017-02-02,Gliclazide 80mg tablets,6.1.2.1
2017-01-10,Gliclazide 80mg tablets,6.1.2.1
2016-11-16,Gliclazide 80mg tablets,6.1.2.1
2017-07-19,Gliclazide 80mg tablets,6.1.2.1
2017-02-24,Humulin M3 KwikPen 100units/ml suspension for injection 3ml pre-filled pen (Eli Lilly and Company Ltd),6.1.1.51
2017-01-04,Metformin 500mg tablets,6.1.2.2
2017-01-04,Humulin M3 KwikPen 100units/ml suspension for injection 3ml pre-filled pen (Eli Lilly and Company Ltd),6.1.1.51
2016-11-14,Metformin 500mg tablets,6.1.2.2
2016-12-05,Humulin M3 KwikPen 100units/ml suspension for injection 3ml pre-filled pen (Eli Lilly and Company Ltd),6.1.1.51
2017-02-28,Metformin 500mg tablets,6.1.2.2
2017-01-30,Metformin 500mg tablets,6.1.2.2
2017-04-24,Metformin 500mg tablets,6.1.2.2
2017-04-24,Metformin 500mg tablets,6.1.2.2
2017-04-10,Humulin M3 KwikPen 100units/ml suspension for injection 3...,6.1.1.51
2017-03-27,Metformin 500mg tablets,6.1.2.2
2017-02-28,Metformin 500mg tablets,6.1.2.2
2017-02-24,Humulin M3 KwikPen 100units/ml suspension for injection 3...,6.1.1.51
2017-01-30,Metformin 500mg tablets,6.1.2.2
2017-01-04,Metformin 500mg tablets,6.1.2.2
2017-01-04,Humulin M3 KwikPen 100units/ml suspension for injection 3...,6.1.1.51
2016-12-08,Metformin 500mg tablets,6.1.2.2
2016-12-05,Humulin M3 KwikPen 100units/ml suspension for injection 3...,6.1.1.51
2017-03-27,Metformin 500mg tablets,6.1.2.2
2017-03-27,FreeStyle Optium testing strips (Abbott Laboratories Ltd),6.1.6.0
,2017-06-21,Metformin 500mg tablets,6.1.2.2
,2017-06-21,Gliclazide 80mg tablets,6.1.2.1
,2017-03-24,Gliclazide 80mg tablets,6.1.2.1
,2017-03-24,Metformin 500mg tablets,6.1.2.2
,2017-02-21,Gliclazide 80mg tablets,6.1.2.1
,2017-06-14,Pioglitazone 30mg tablets,6.1.2.3
,2017-04-28,Vildagliptin 50mg / Metformin 850mg tablets,6.1.2.3
,2017-04-07,Metformin 500mg tablets,6.1.2.2
,2017-04-07,NovoMix 30 FlexPen 100units/ml suspension for injection 3...,6.1.1.51
,2017-03-09,Sitagliptin 100mg tablets,6.1.2.3
,2017-04-24,Gliclazide 80mg tablets,6.1.2.1
,2017-01-30,NovoRapid Penfill 100units/ml solution for injection 3ml ...,6.1.1.1
,2017-01-06,Gliclazide 80mg tablets,6.1.2.1
,2016-12-16,GlucaGen Hypokit 1mg powder and solvent for solution for ...,6.1.4.0
,2016-12-15,GlucoGel 40% gel original (BBI Healthcare Ltd),6.1.4.0
,2016-11-22,Gliclazide 80mg tablets,6.1.2.1
,2017-01-06,Metformin 850mg tablets,6.1.2.2
,2017-04-28,Pioglitazone 30mg tablets,6.1.2.3
,2017-02-28,Glipizide 5mg tablets,6.1.2.1
,2017-09-27,Glipizide 5mg tablets,6.1.2.1
,2017-01-09,GlucoGel 40% gel original (BBI Healthcare Ltd),6.1.4.0
,2017-03-07,Metformin 850mg tablets,6.1.2.2
,2016-11-14,FreeStyle Optium testing strips (Abbott Laboratories Ltd),6.1.6.0
,2016-12-08,Metformin 500mg tablets,6.1.2.2
,2017-04-10,Humulin M3 KwikPen 100units/ml suspension for injection 3ml pre-filled pen (Eli Lilly and Company Ltd),6.1.1.51
,2017-06-21,Gliclazide 80mg tablets,6.1.2.1
,2016-11-14,Metformin 500mg tablets,6.1.2.2
,2017-05-22,Gliclazide 80mg tablets,6.1.2.1
,2017-04-24,Gliclazide 80mg tablets,6.1.2.1
,2017-04-24,Metformin 500mg tablets,6.1.2.2
,2017-01-26,Metformin 500mg tablets,6.1.2.2
,2017-01-26,Gliclazide 80mg tablets,6.1.2.1
,2016-12-21,Metformin 500mg tablets,6.1.2.2
,2016-12-21,Gliclazide 80mg tablets,6.1.2.1
,2016-12-19,Metformin 500mg tablets,6.1.2.2
,2017-10-24,Pioglitazone 30mg tablets,6.1.2.3
,2017-10-24,Vildagliptin 50mg / Metformin 850mg tablets,6.1.2.3
,2017-08-14,Vildagliptin 50mg / Metformin 850mg tablets,6.1.2.3
,2017-08-14,Pioglitazone 30mg tablets,6.1.2.3
,2017-03-03,Vildagliptin 50mg / Metformin 850mg tablets,6.1.2.3
,2017-03-03,Pioglitazone 30mg tablets,6.1.2.3
,2017-01-11,Pioglitazone 30mg tablets,6.1.2.3
,2016-11-12,Vildagliptin 50mg / Metformin 850mg tablets,6.1.2.3
,2017-06-14,Vildagliptin 50mg / Metformin 850mg tablets,6.1.2.3
,2016-11-08,Pioglitazone 30mg tablets,6.1.2.3
,2016-12-30,Vildagliptin 50mg / Metformin 850mg tablets,6.1.2.3
,2017-05-17,NovoRapid Penfill 100units/ml solution for injection 3ml cartridges (Novo Nordisk Ltd),6.1.1.1
,2017-05-17,Levemir Penfill 100units/ml solution for injection 3ml cartridges (Novo Nordisk Ltd),6.1.1.2
,2017-05-17,NovoRapid Penfill 100units/ml solution for injection 3ml ...,6.1.1.1
,2017-05-17,Levemir Penfill 100units/ml solution for injection 3ml ca...,6.1.1.2
,2017-03-08,NovoRapid Penfill 100units/ml solution for injection 3ml ...,6.1.1.1
,2016-12-22,NovoRapid Penfill 100units/ml solution for injection 3ml ...,6.1.1.1
,2016-12-22,Levemir Penfill 100units/ml solution for injection 3ml ca...,6.1.1.2
,2017-03-08,NovoRapid Penfill 100units/ml solution for injection 3ml cartridges (Novo Nordisk Ltd),6.1.1.1
,2017-03-08,Levemir Penfill 100units/ml solution for injection 3ml cartridges (Novo Nordisk Ltd),6.1.1.2
,2016-12-22,NovoRapid Penfill 100units/ml solution for injection 3ml cartridges (Novo Nordisk Ltd),6.1.1.1
,2016-12-22,Levemir Penfill 100units/ml solution for injection 3ml cartridges (Novo Nordisk Ltd),6.1.1.2
,2017-06-15,NovoMix 30 FlexPen 100units/ml suspension for injection 3...,6.1.1.51
,2017-06-15,Metformin 500mg tablets,6.1.2.2
,2017-01-25,Sitagliptin 100mg tablets,6.1.2.3
,2017-01-24,Glipizide 5mg tablets,6.1.2.1
,2016-12-13,Glipizide 5mg tablets,6.1.2.1
,2016-11-03,Glipizide 5mg tablets,6.1.2.1
```

managing time series data – 3
drug combinations as words - for natural language processing approach



Drug Sentence: MF, GLP1_MF, GLP1_MF_SGLT2, GLP1_MF

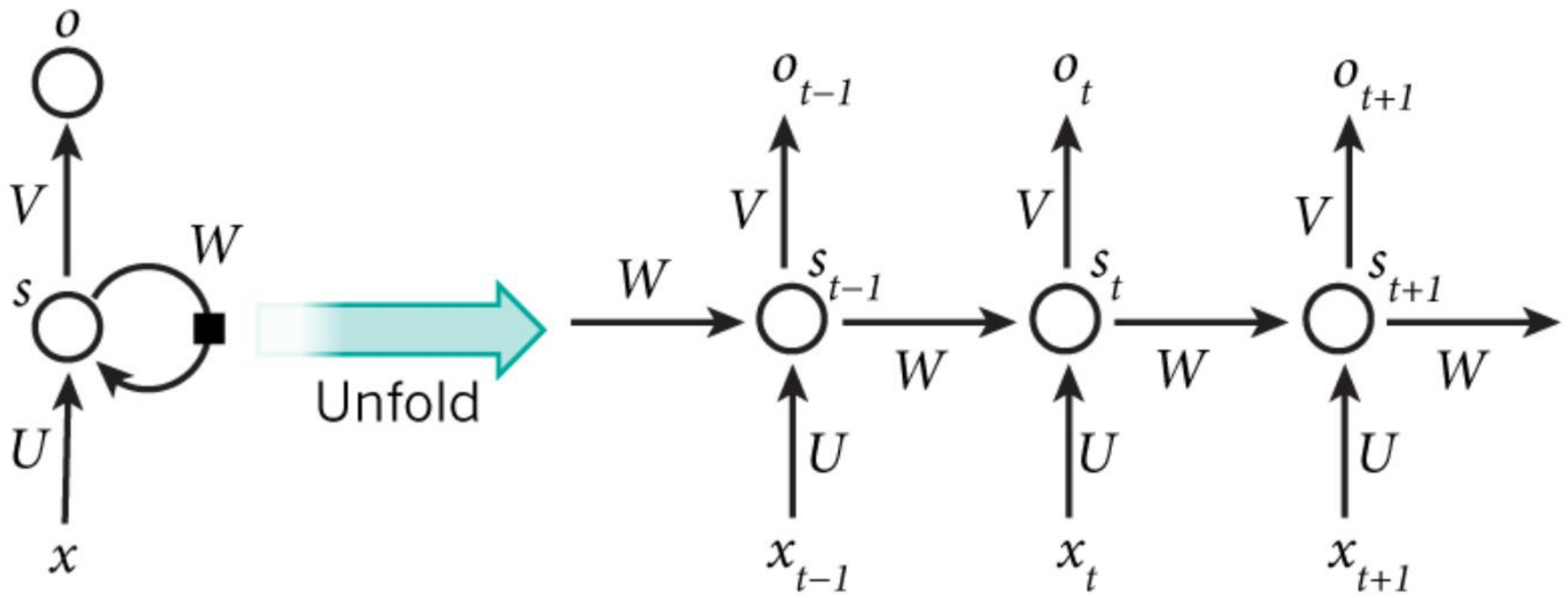
↓

Embedding -> numerical vector

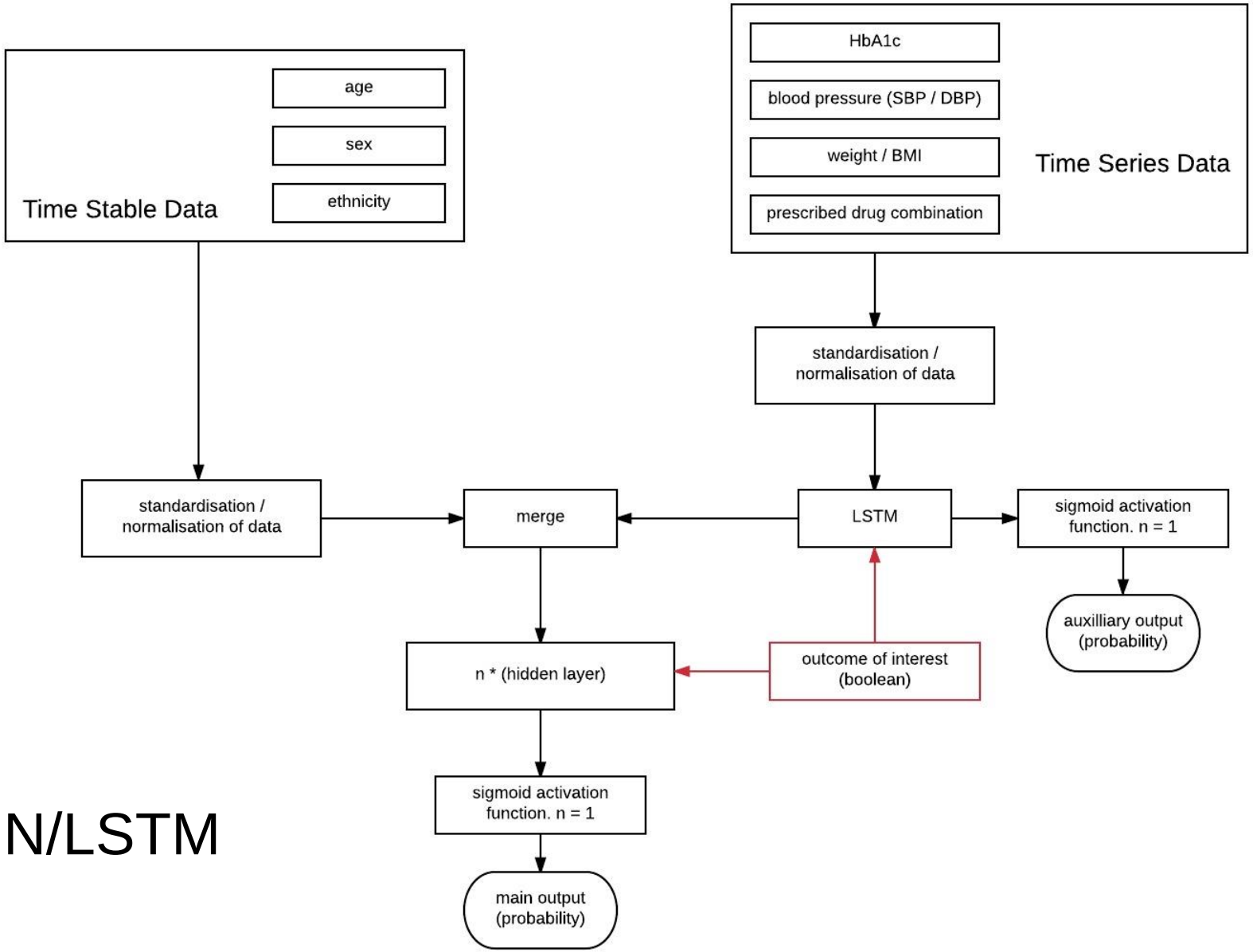
↓

input into RNN / LSTM

Recurrent Neural Network (LSTM)

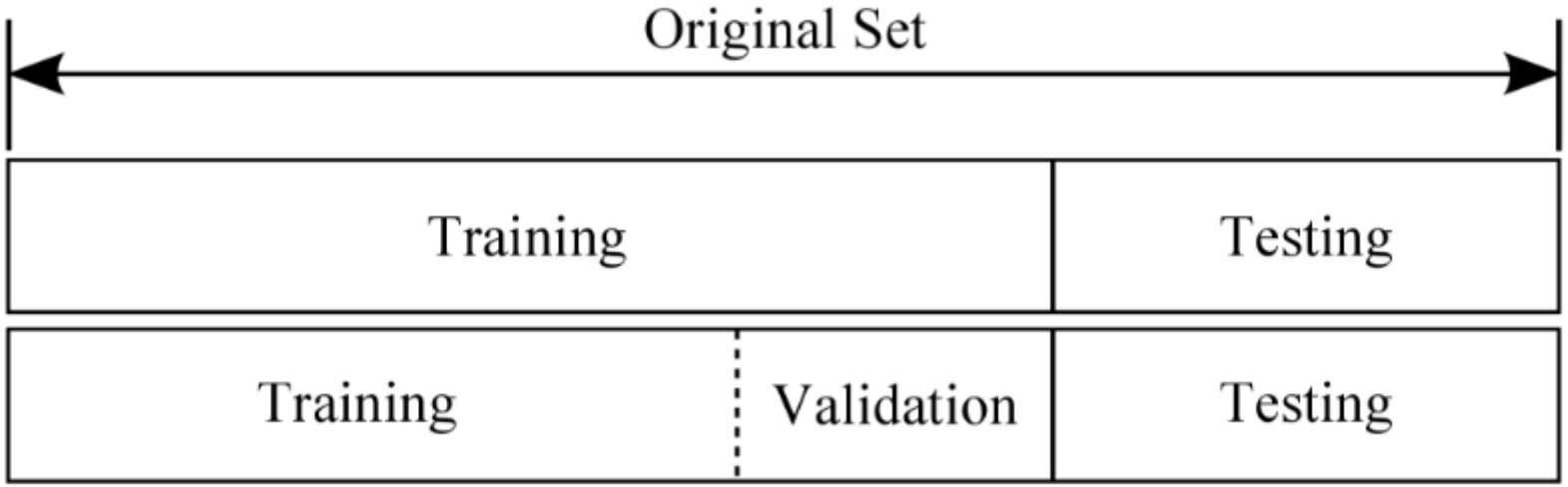


extracts information from sequence of input
multi-dimensional RNN extracts information from the interactions between input sequences over time



schematic of RNN/LSTM based classifier

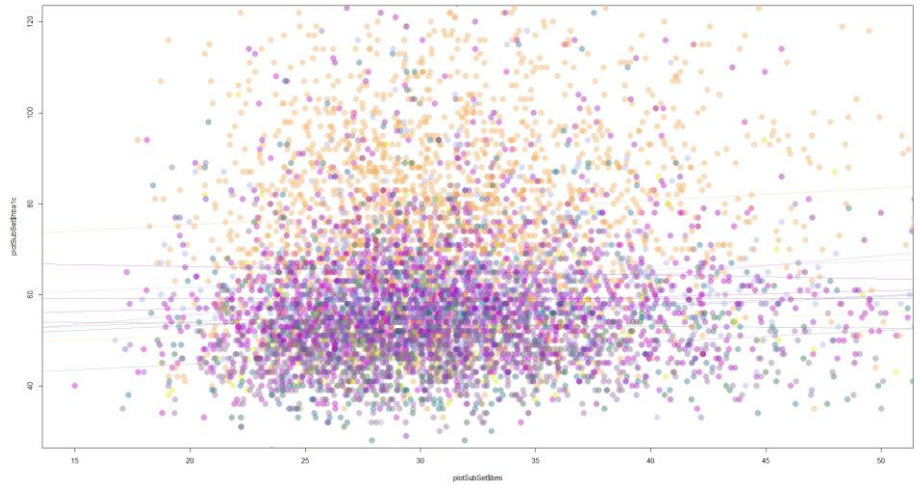
training, validation and withheld test sets



environments used:

SCI diabetes data input

visualisations etc



outcome measure of interest

eg

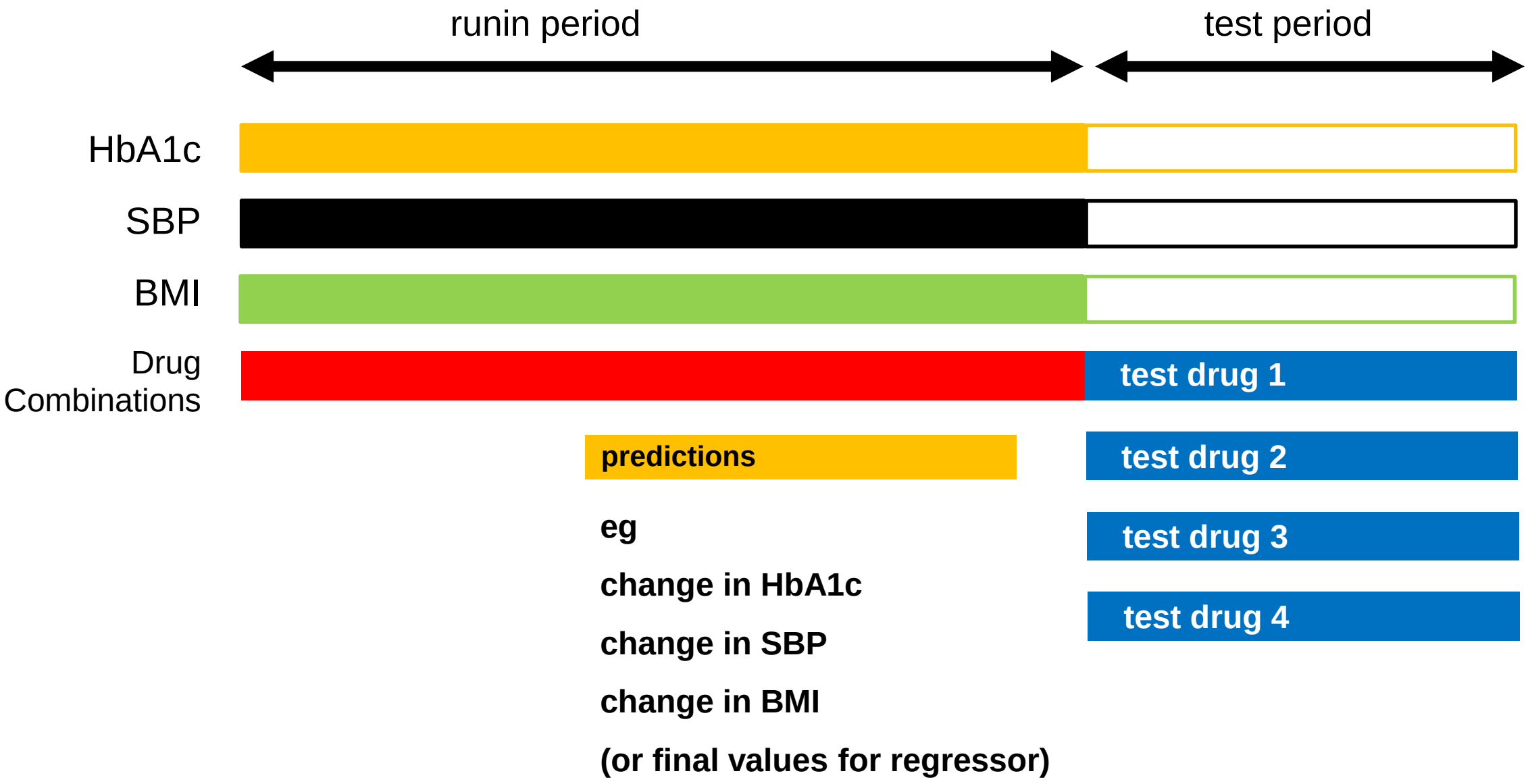
change in HbA1c

change in SBP

change in BMI

(or final values for regressor version)

training data structure



using the model to predict response

bidirectional LSTM classifier

i. simple classification endpoint:

   eg

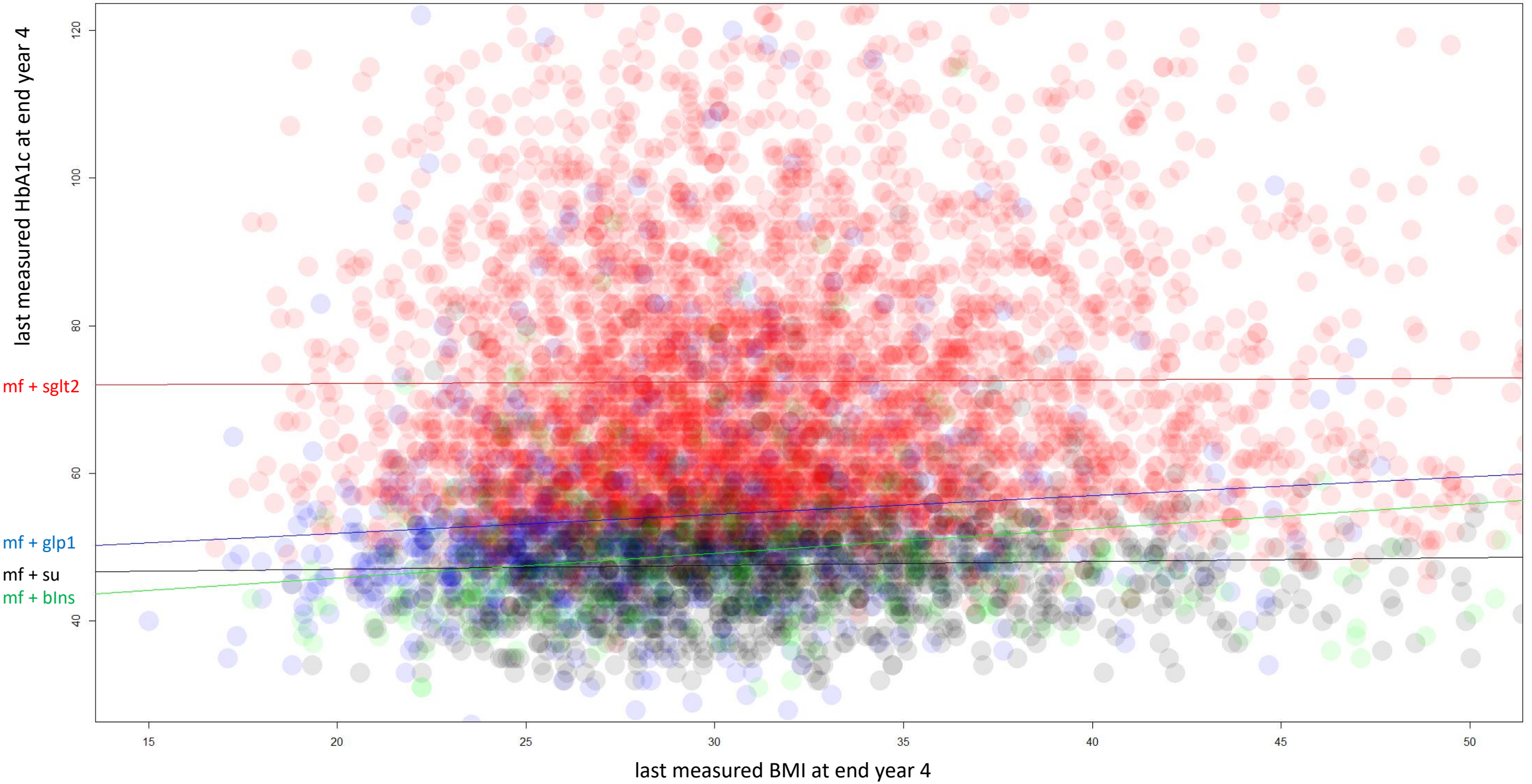
   - probability of a 10mmol/mol reduction in hba1c at 1y
   - probability of achieving HbA1c in range 48 - 60mmol/mol at 1y

ii. composite endpoints:

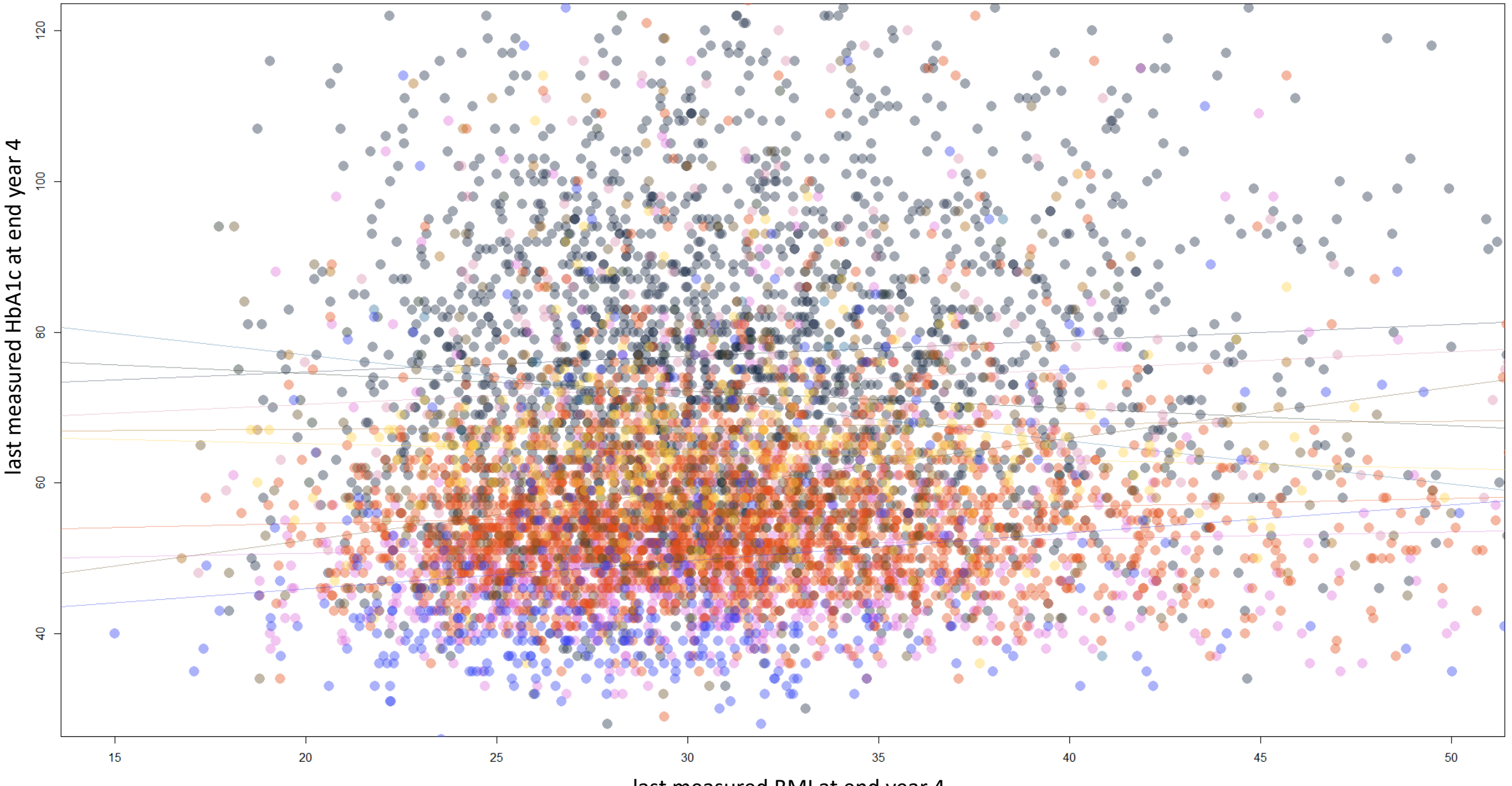
   eg

   probability of achieving HbA1c 48 - 60mmol/mol, with a SBP of <140mmHg

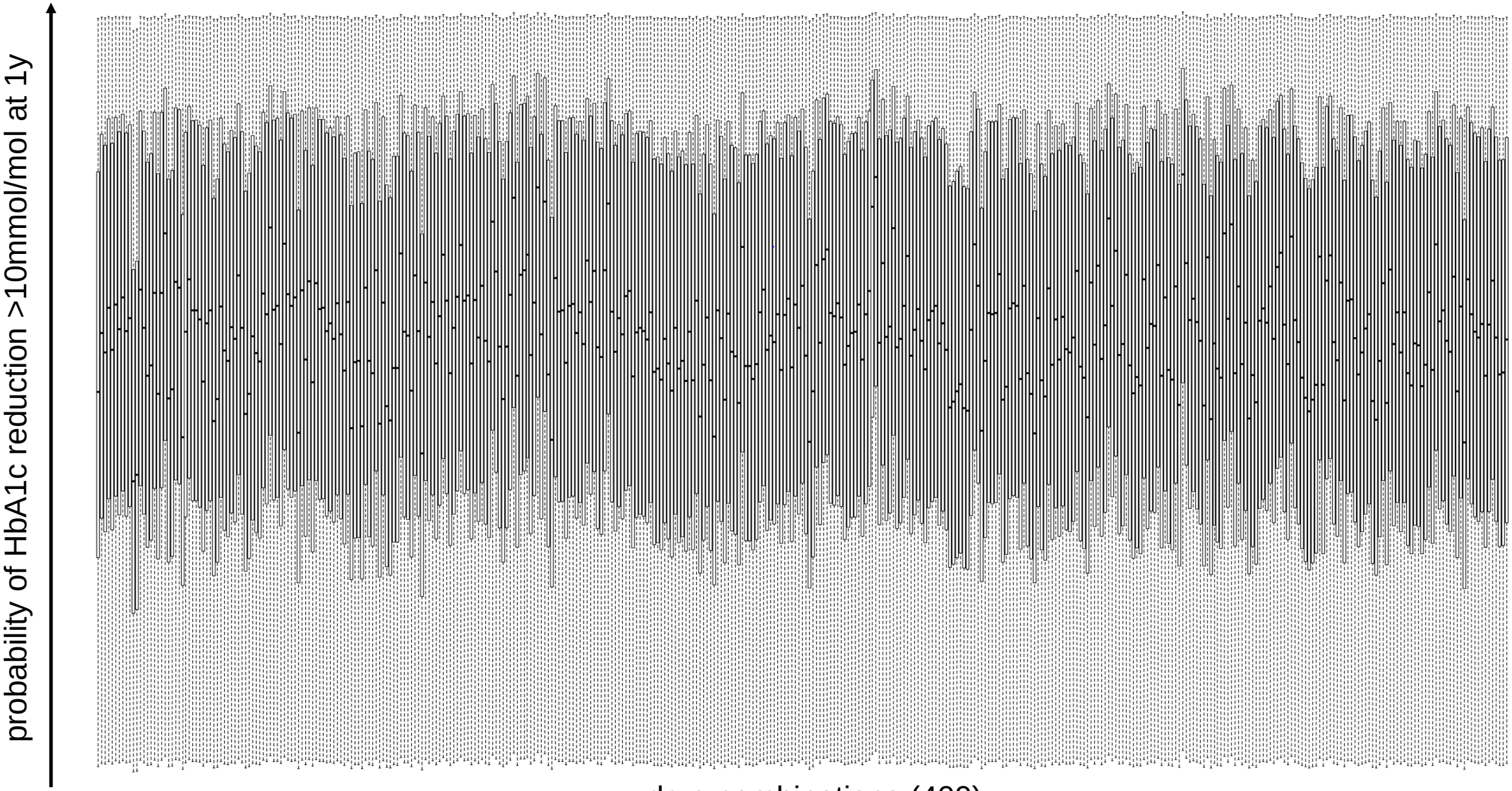
simple problem: which of 4 combinations most likely to reduce hba1c by 10mmol/mol?

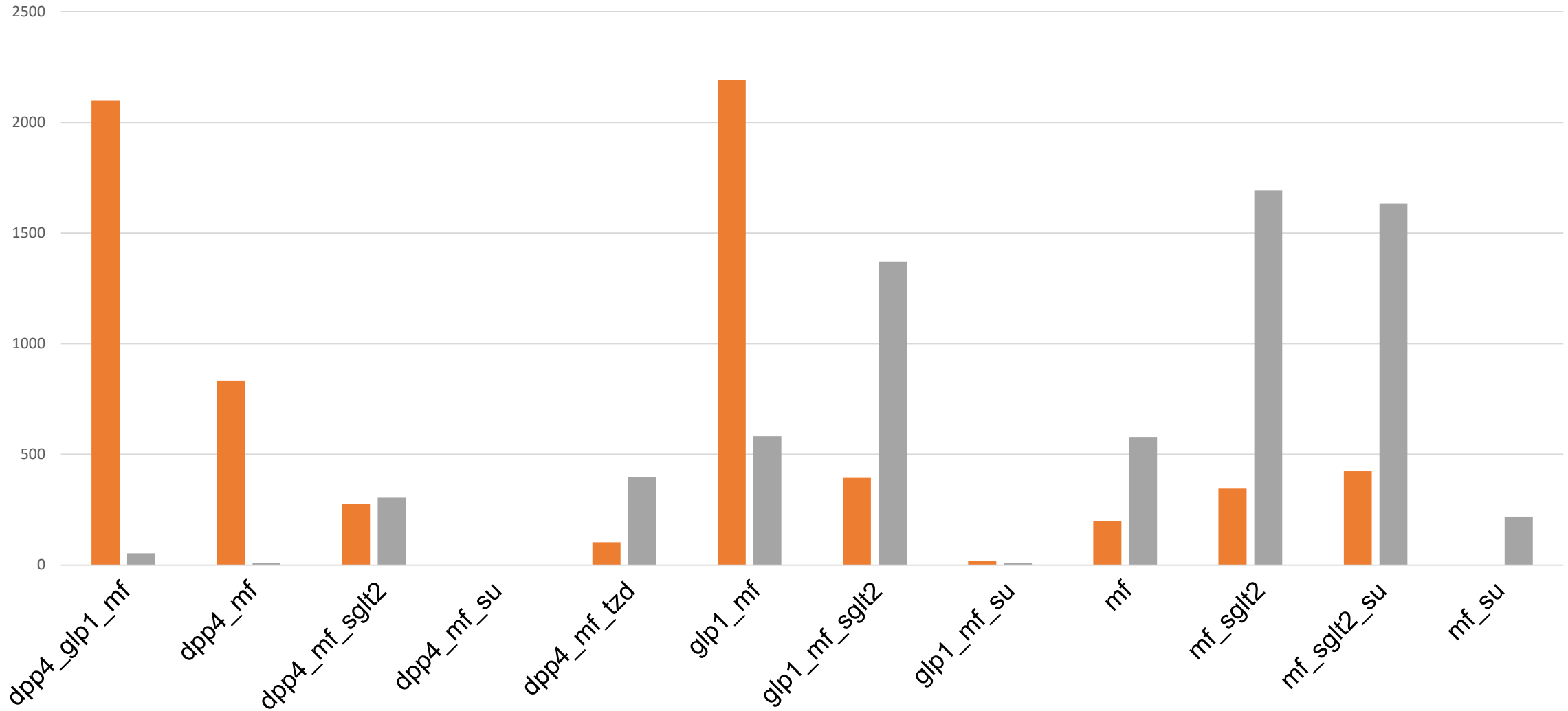


more complex problem: which of 16 combinations most likely to reduce hba1c to <60mmol/mol without causing weight gain?



probability of HbA1c reduction >10mmol/mol at 1y n = 1450 (initial HbA1c >60mmol/mol)

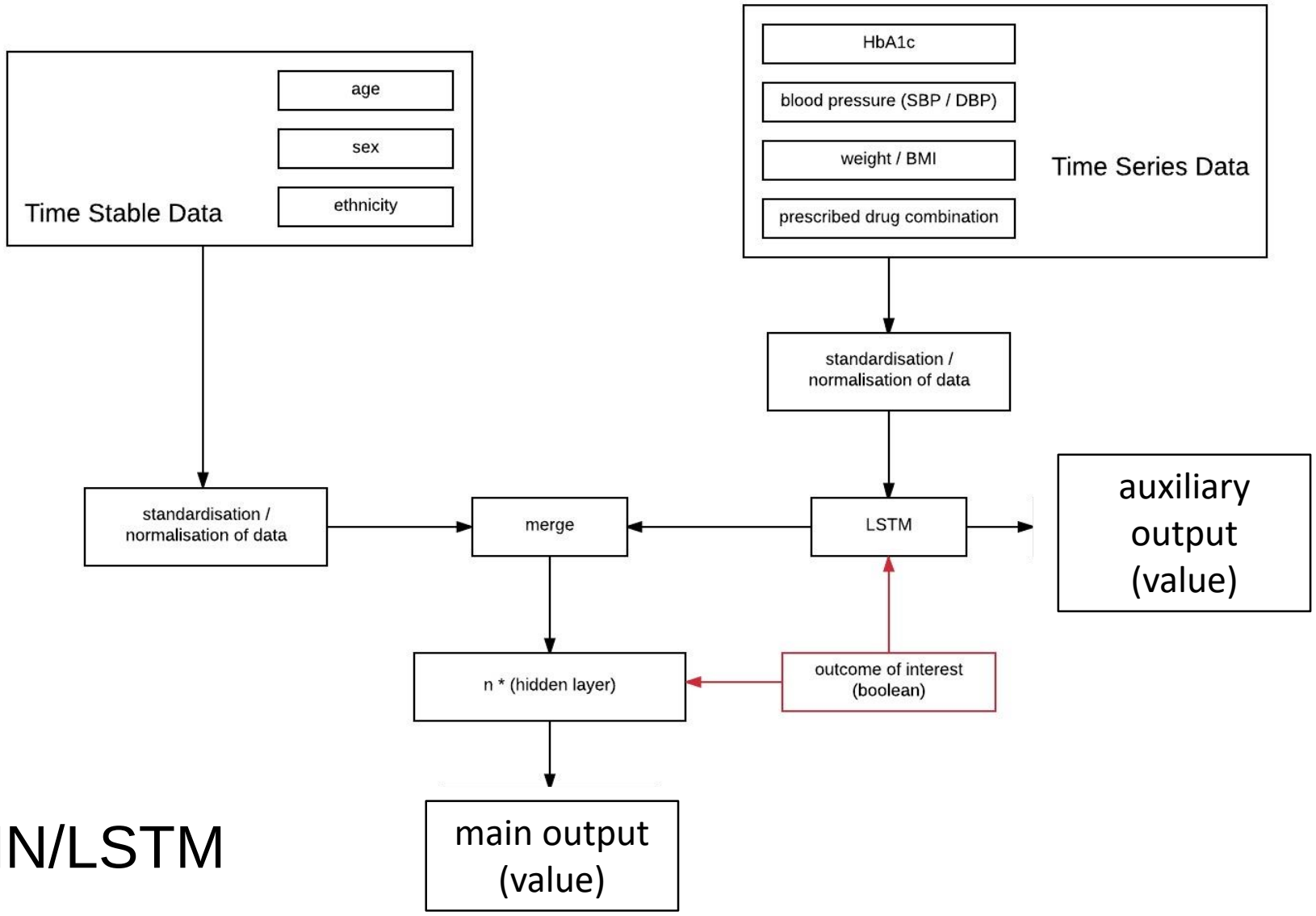



2500
2000
1500
1000
500
0
dpp4_glp1_mf
dpp4_mf
dpp4_mf_sglt2
dpp4_mf_su
dpp4_mf_tzd
glp1_mf
glp1_mf_sglt2
glp1_mf_su
mf
mf_sglt2
mf_sglt2_su
mf_su
best combination to achieve HbA1c <60mmol/mol, with reduction in BMI
best combination to achieve HbA1c <60mmol/mol, with reduction in SBP

regressor

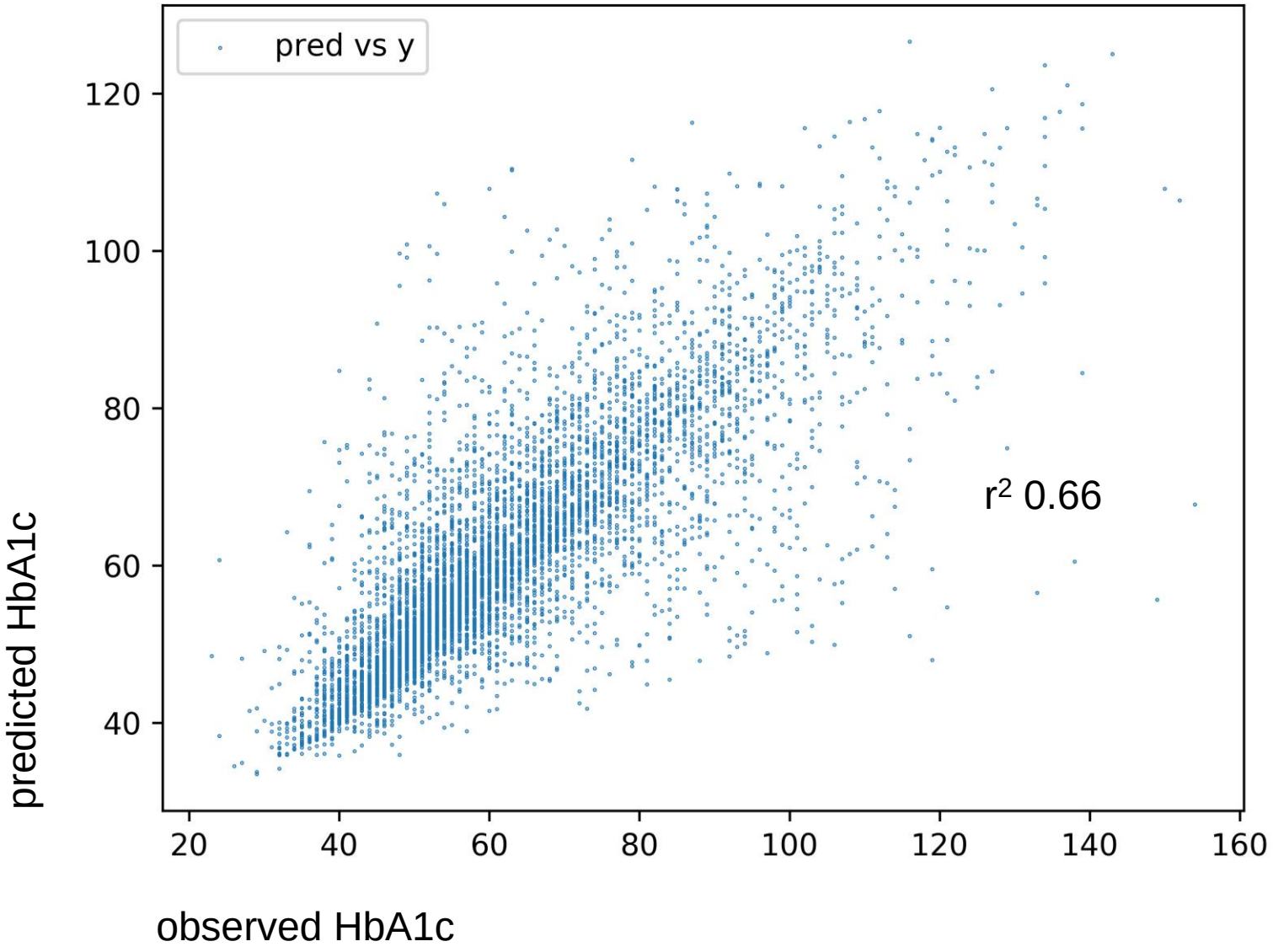
remove sigmoid activating function from final layer

-> returns value that can be mapped to predicted outcome value (eg HbA1c)

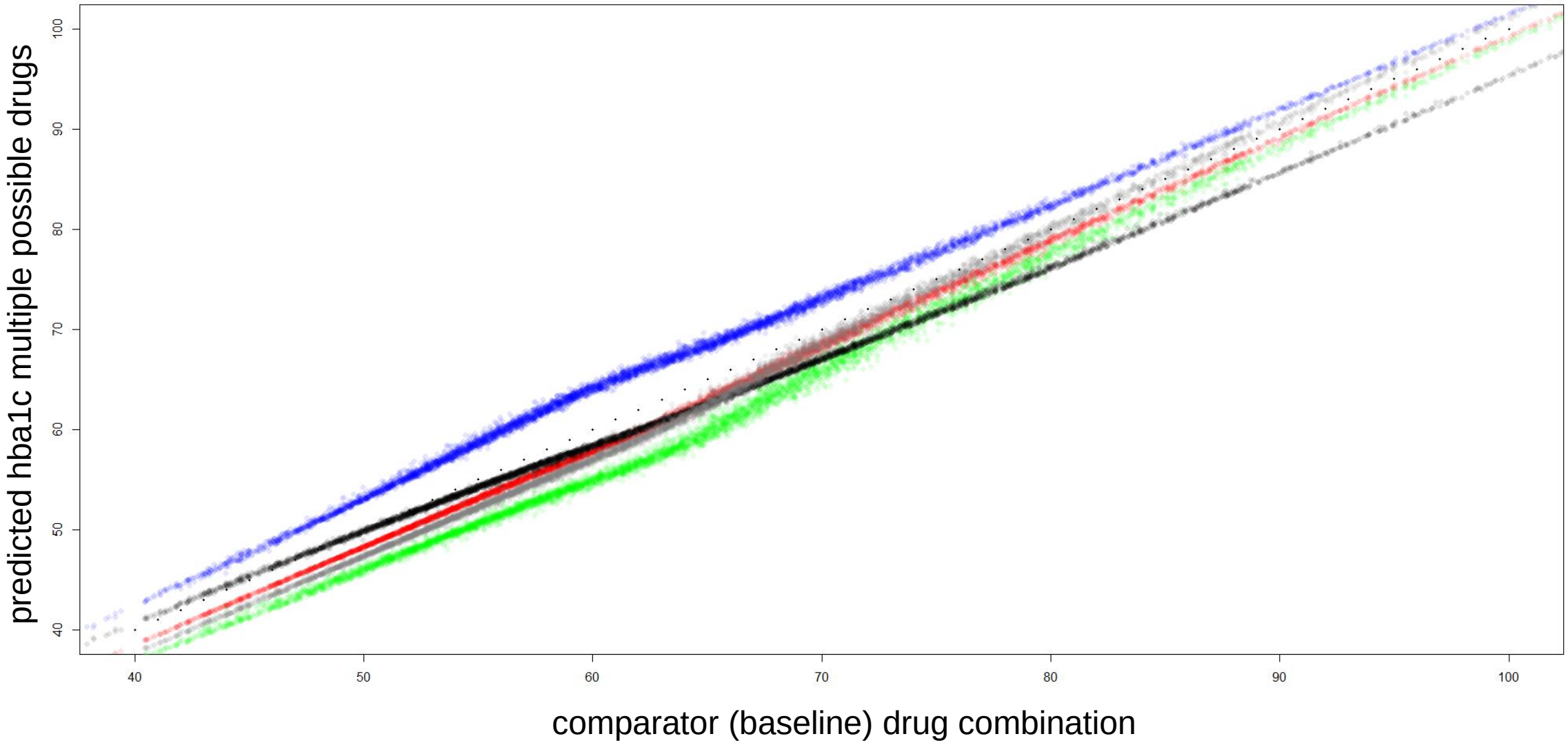


schematic of RNN/LSTM based regressor

model as regressor – prediction of HbA1c at 12 months vs actual



representation of predicted drug effects



current performance

classifier – AUC (modelling actual drug therapy) 0.75 – 0.85

regressor – r^2 ~0.65

ongoing work

optimise neural network / layers used (LSTM vs convolutional etc)

explore alternative techniques (Gokhale/Tino baysian inference approach etc)

some current issues

imbalanced classes	downsampling / upsampling approaches
stochastic variation	k-fold validation / averaging multiple runs
hyperparameter tuning	increase computational power / use small samples
diabetes classification issues	unsupervised reclassification
data sharing	synthetic data generation

diabetes classification issues

MSc (Stratified Medicine) project 2017

Machine-Learning (neural network) driven algorithmic classification of Type 1 or Type 2 diabetes at the time of presentation significantly outperforms experienced clinician classification

CAR Sainsbury, R Muir, J Osmanska, GC Jones

glucose.ai Research Group, Department of Diabetes, Gartnavel General Hospital, Glasgow, Scotland

NHS Greater Glasgow and Clyde

http://glucose.ai
@csainsbury
c.sainsbury@nhs.net

Background / Aims

Classification of type of diabetes at the point of diagnosis may not always be straightforward. We aimed to develop an algorithmic approach to the problem, using data available within the SCI-Diabetes dataset. Metrics that would be routinely available at the time of first presentation were chosen for inclusion. To asses the potential clinical utility - and to benchmark performance - a subset of individuals were presented to experienced clinicians who classified as either Type 1 or Type 2 diabetes. The accuracy of classifications generated for individuals within the subset by both algorithm and by clinicians were compared, using the established diagnosis within SCI diabetes as the comparator diagnosis. Individuals were included in the analysis only if they had a date of diagnosis at least 12 months from the data extraction date, in order to ensure a high likelihood of a stable and correct diagnosis being achieved.

Data was prepared for analysis using R(1), and analysis code was written in Python(2). An artificial neural network was chosen for the algorithmic approach, implemented in Tensorflow (3) (written using the Keras library (4))

A further subset of the test subset was generated for clinician classification. Data from individuals in this subset were presented in batches of 100 to experienced clinicians within our clinic. The proportion of individuals with a recorded diagnosis of Type 1 Diabetes within each batch was varied at random between 0.05 and 0.5 to reduce the possibility of clinicians inferring a diagnosis from the proportion already identified.

The ANN model was applied to the test subset, with the output being a probability for each individual of the correct diagnosis being Type 1 Diabetes. A Receiver Operator Characteristic (ROC) curve was generated using these probabilities. For Sensitivity / Specificity analyses a threshold of 0.2 was applied to the probability value – if the probability was abo this threshold a diagnosis of Type 1 Diabetes was deemed to be predicted.

In the case of the subset of patients presented to clinicians fo classification, a confusion matrix was generated for both algorithmic and clinician classification outcomes, and Sensitivity, Specificity, Positiv Predictive Value and Negative Predictive value were calculated.

METHODS

- **SCI-Diabetes NHS Greater Glasgow and Clyde data refinement**

- Type 1 and Type 2 patients only

- **Parameters selected:**

- BMI, Systolic BP, Diastolic BP, HbA1c levels, Age, Gender and Ethnicity.

Table 1: Patient Information Sheet Example

Ethnicity	BMI	SBP	DBP	HbA1c	Age	Gender
White - Scottish	25.83	118	76	84	78.49	male

Table 1 illustrates an example of the information sheet provided to physicians. The same information was then programmed into both the logistic regression and ANN. The information was given as shown to ensure no unfair bias was given to any model during diagnosis.

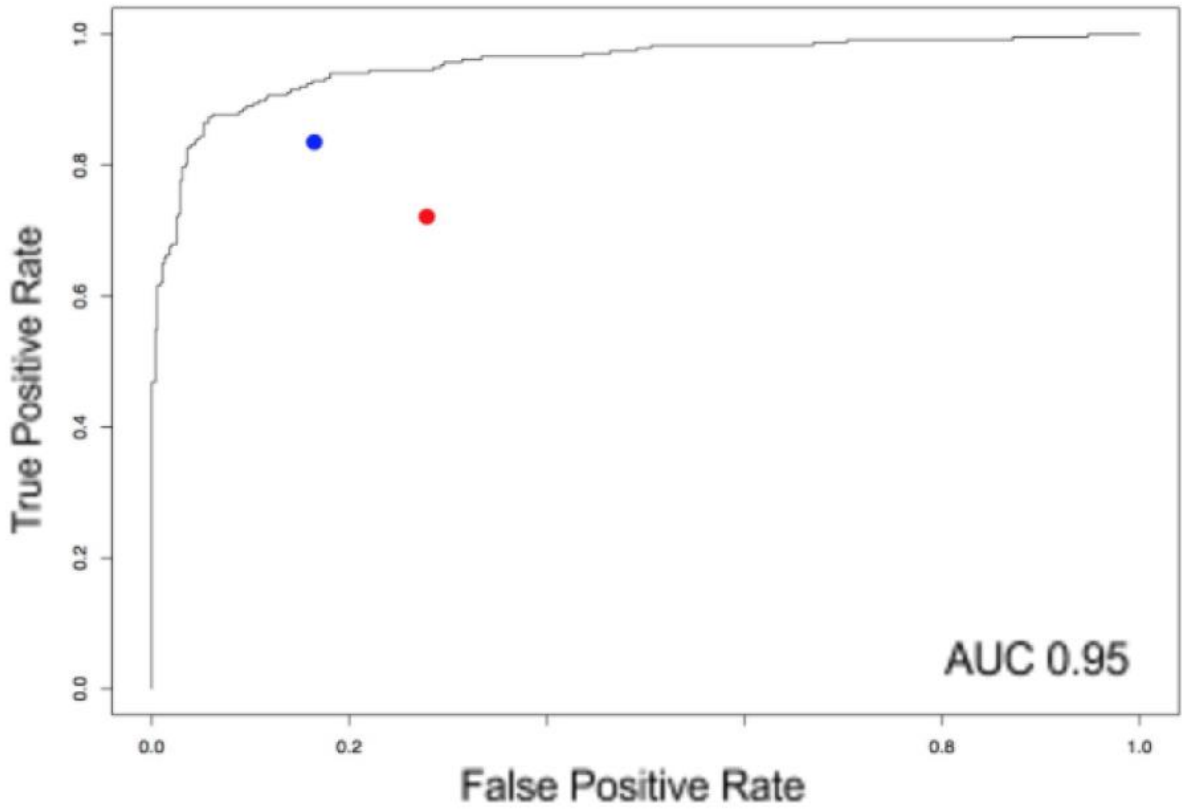
RESULTS

Table 1. Confusion Matrix Results

	Physician	ANN	Logistic Regression
Accuracy (CI 0.95)	0.86 (0.78,0.92)	0.93 (0.85, 0.96)	0.91 (0.83, 0.96)
Specificity	0.77	0.85	0.81
Sensitivity	0.93	0.96	0.95

Table 1 illustrates the median confusion matrix values collected from the 6 physician forms analysed. The table details accuracy(CI 0.95), specificity and sensitivity for each model which was then utilised in the development of the ROC curve analysis.

Figure 2. ROC curve analysis



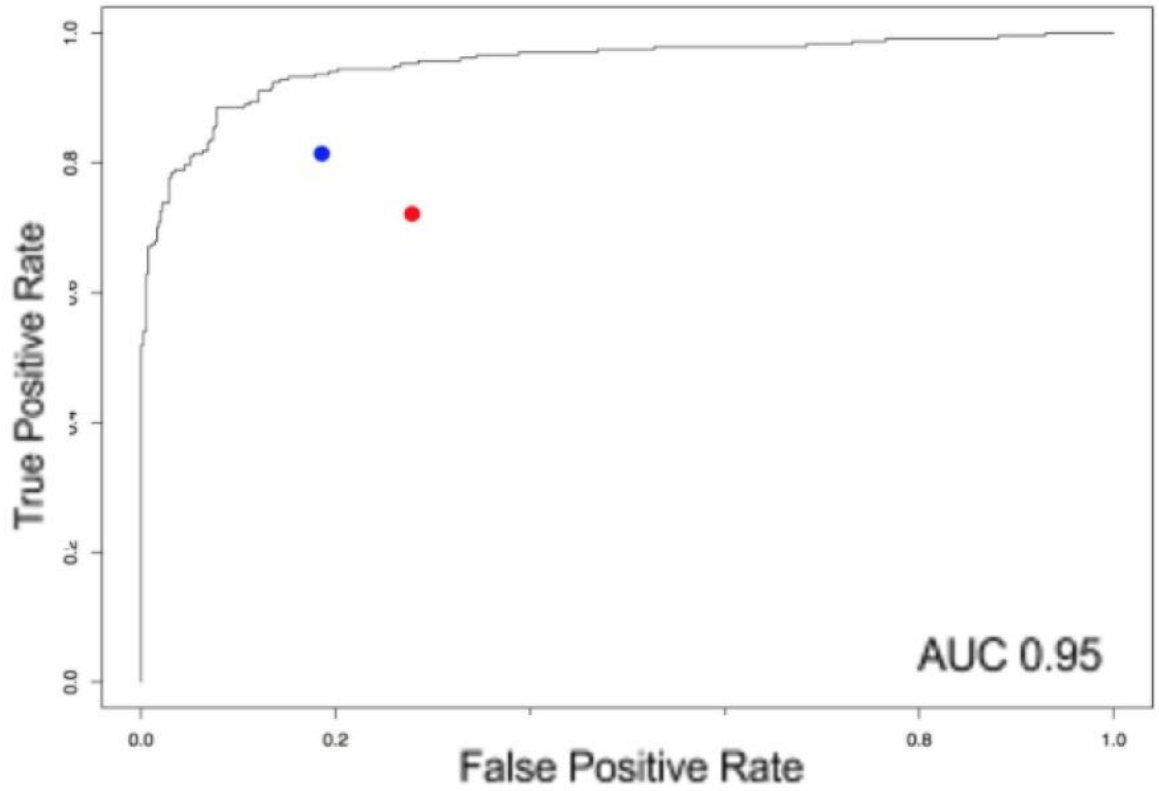


Figure 2 illustrates a full dataset analysis of the ANN model (left) and the logistic regression (right). On each model relevant performance of the physician(red dot) versus the model(blue dot) are also plotted to illustrate model superiority.

Figure 3. ANN and patient characteristic association

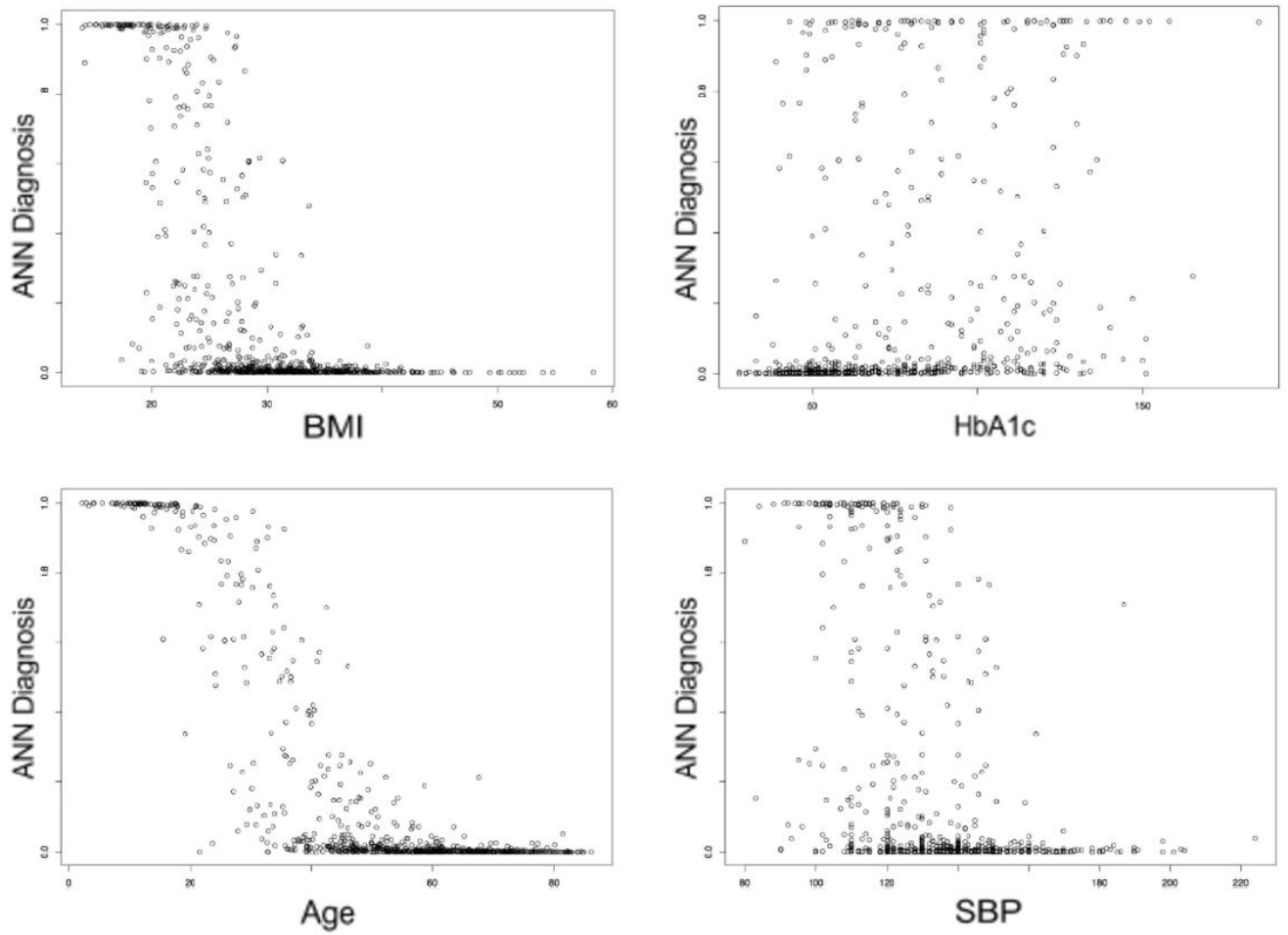


Figure 3 illustrates the association between examined patient characteristics and ANN model predicted diagnosis. BMI (top left), HbA1c (top right), Age (bottom left) and SBP (bottom right) are illustrated.

Table 2. Statistical Summary

Factors	General coh ort	Type 1 patients	Type 2 patients
Cohort no.	49,995	3,222	46,773
Sex	F: 22,124 M: 27,871	F: 1,388 M: 1,834	F: 20,736 M: 26,037
Mean Age (years)	56.7 (48.0, 67.2)	28.5 (15.1, 39.9)	58.6 (49.8, 67.8)
Mean BMI	31.8 (27.2, 35.6)	23.4 (19.5, 26.4)	32.4 (27.8, 35.9)
Mean hba1c	66.0 (49.0, 79.0)	86.1 (63.0, 107.0)	64.9 (49.0, 77.0)
Mean sbp	136.5 (124.0, 147.0)	120.8 (110.0, 130.0)	137.5 (125.0, 148.0)
Mean dbp	80.2 (72.0, 87.0)	72.6 (63.0, 80.0)	80.7 (73.0, 88.0)

problem

are labels accurate?

Table 2. Statistical S[illegible]mary

Factors	General [illegible]ort	Type 1 patients	Type 2 patients
Cohort no.	49,995	3,222	46,773
Sex	F: 22,124 M: 27,871	F: 1,388 M: 1,834	F: 20,736 M: 26,037
Mean Age (years)	56.7 (48.0, 67.2)	28.5 (15.1, 39.9)	58.6 (49.8, 67.8)
Mean BMI	31.8 (27.2, 35.6)	23.4 (19.5, 26.4)	32.4 (27.8, 35.9)
Mean hba1c	66.0 (49.0, 79.0)	86.1 (63.0, 107.0)	64.9 (49.0, 77.0)
Mean sbp	136.5 (124.0, 147.0)	120.8 (110.0, 130.0)	137.5 (125.0, 148.0)
Mean dbp	80.2 (72.0, 87.0)	72.6 (63.0, 80.0)	80.7 (73.0, 88.0)

problem / opportunities

1
to clean existing data

2
to investigate a data-driven, time series based classification

3
to find ‘missing’ diagnoses (MODY etc)

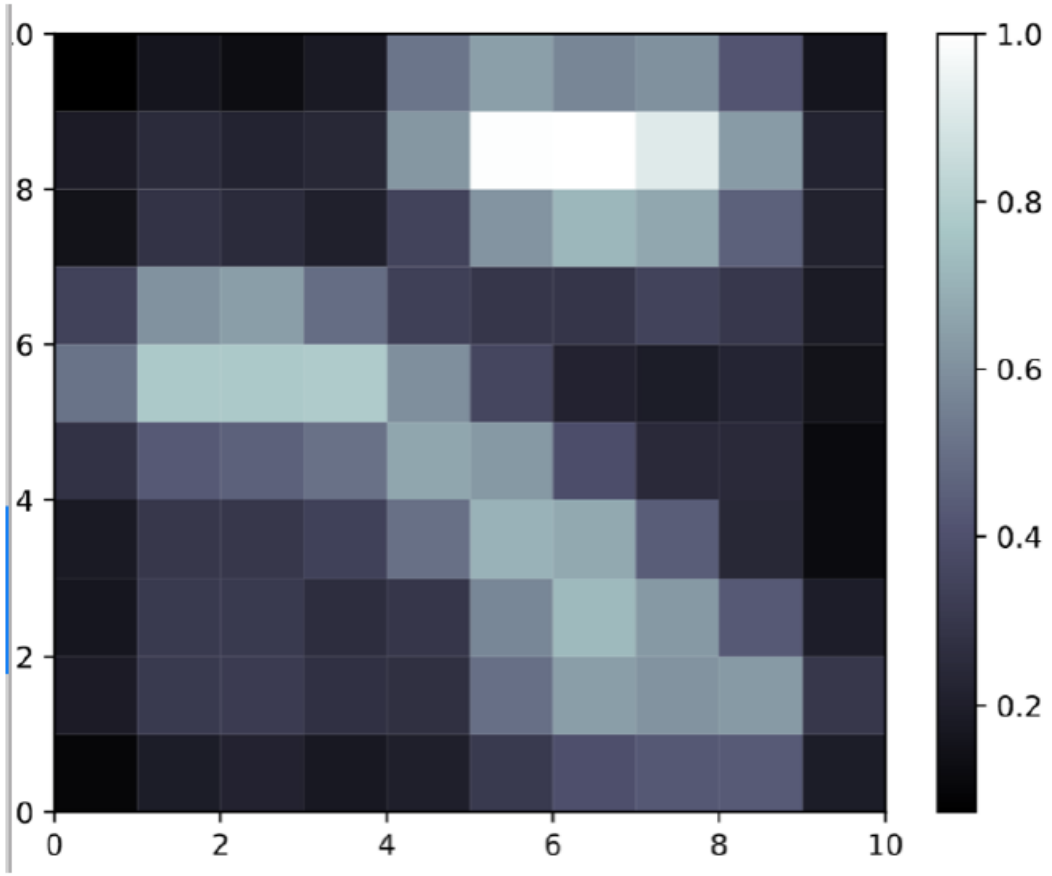
unsupervised approach

self organising map

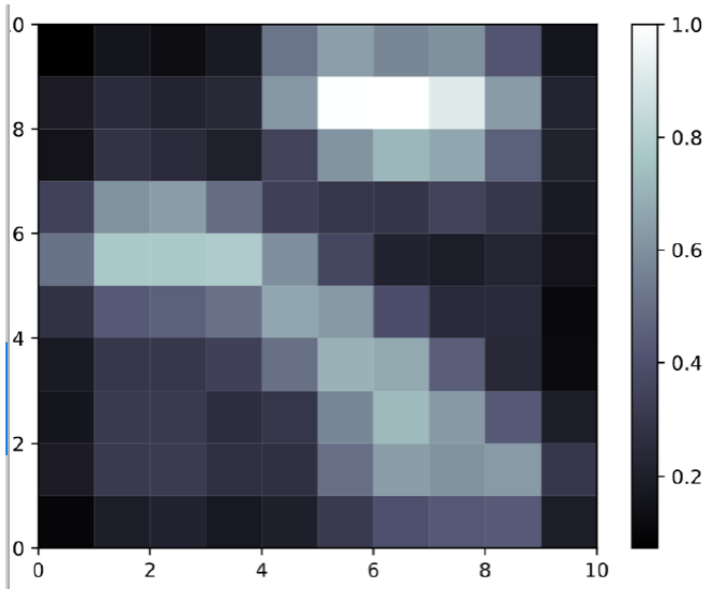
ANN that allows dimensionality reduction

eg
2 dimensional representation of multidimensional input

in this context – can make assessment of probability of the correct label being applied

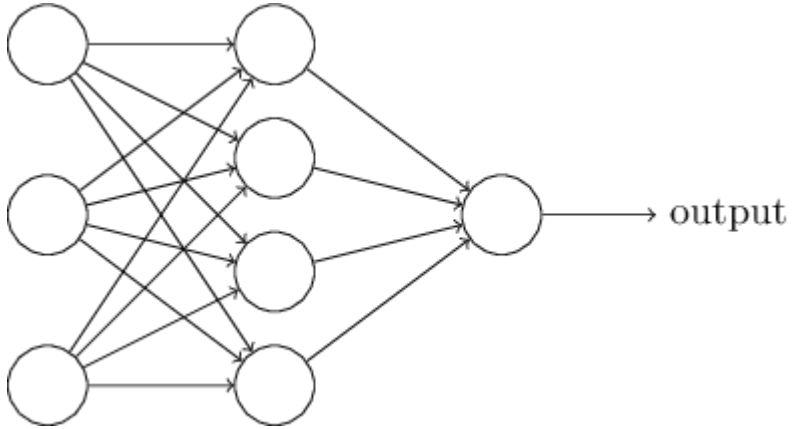


clean existing data

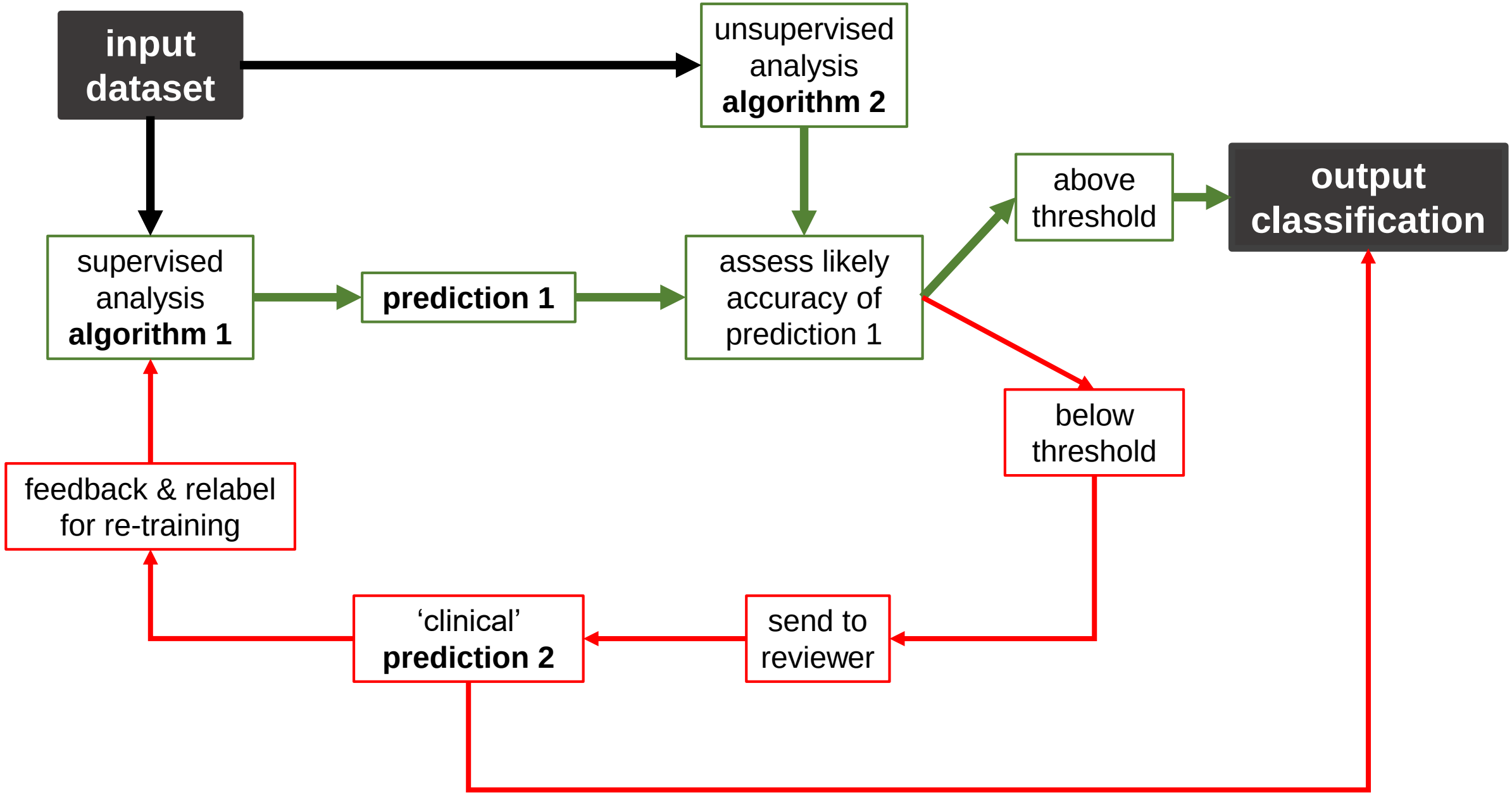


combined supervised / unsupervised approaches

+add human into the loop



clean existing data


input dataset
unsupervised analysis **algorithm 2**
supervised analysis **algorithm 1**
prediction 1
assess likely accuracy of prediction 1
above threshold
output classification
below threshold
feedback & relabel for re-training
'clinical' **prediction 2**
send to reviewer
clean existing data

data-driven, time series based classification

Novel subgroups of adult-onset diabetes and their association with outcomes: a data-driven cluster analysis of six variables

Emma Ahlqvist, Petter Storm, Annemari Käräjämäki, Mats Martinell*, Mozhgan Dorkhan, Annelie Carlsson, Petter Vikman, Rashmi B Prasad, Dina Mansour Aly, Peter Almgren, Ylva Wessman, Nael Shaat, Peter Spégel, Hindrik Mulder, Eero Lindholm, Olle Melander, Ola Hansson, Ulf Malmqvist, Åke Lernmark, Kaj Lahti, Tom Forsén, Tiinamaija Tuomi, Anders H Rosengren, Leif Groop*

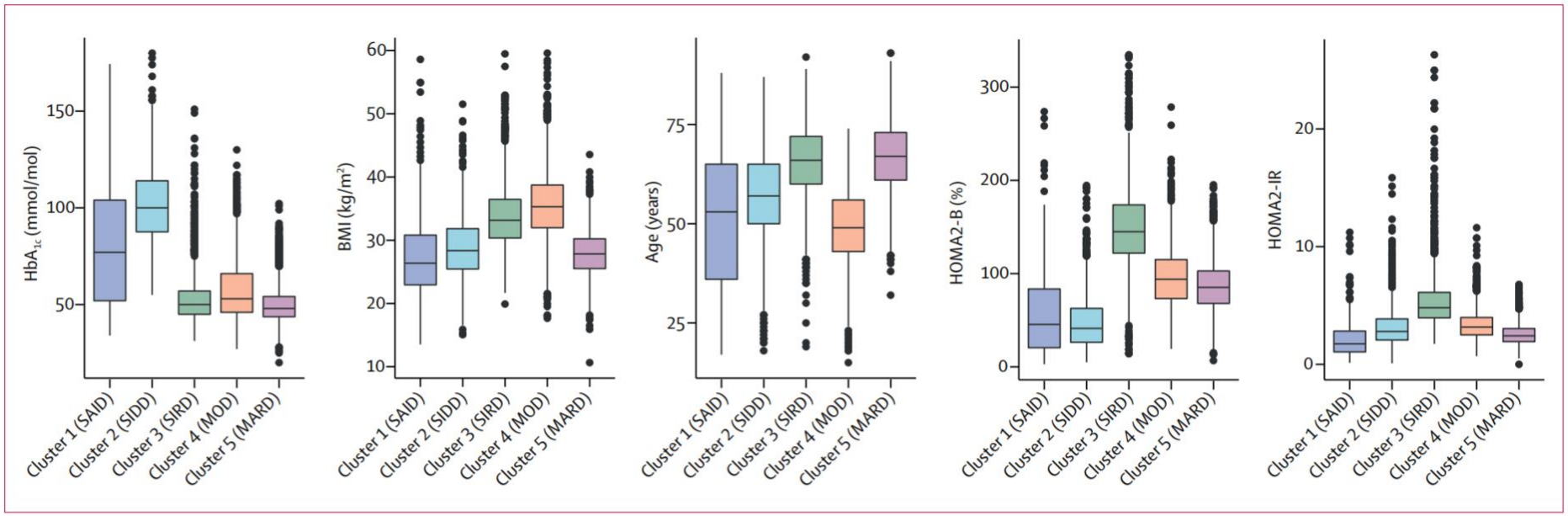
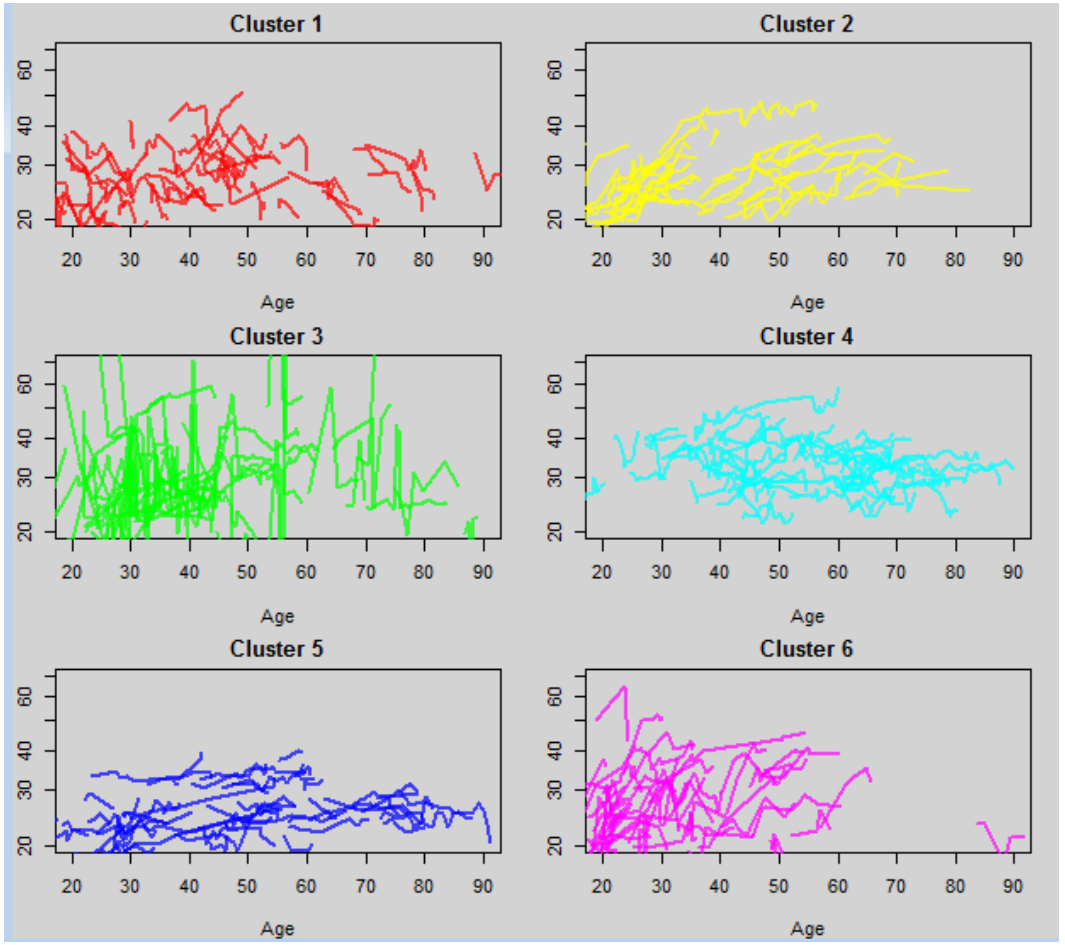


Figure 2: **Cluster characteristics in the ANDIS cohort**
Distributions of HbA_{1c} and age at diagnosis, and BMI, HOMA2-B, and HOMA2-IR at registration, in the ANDIS cohort for each cluster. k-means clustering was done separately for men and women; pooled data are shown here for clusters 2–5. SAID=severe autoimmune diabetes. SIDD=severe insulin-deficient diabetes. SIRD=severe insulin-resistant diabetes. MOD=mild obesity-related diabetes. MARD=mild age-related diabetes. HOMA2-B=homoeostatic model assessment 2 estimates of β-cell function. HOMA2-IR=homoeostatic model assessment 2 estimates of insulin resistance. ANDIS=All New Diabetics in Scania.

2. clustering approach using TS data

- HbA1c
- BP
- BMI
- biochem
- drugs / drug response



Example: pattern mining for BMI trajectory analysis (THIN data)

3. identify missing diagnoses

difficult as small sample size to work from (but may be confident in label accuracy)

upsampling ideal to maximise value from data

synthetic data

potential to help with another problem – data sharing

requirements for synthetic data generator

- generate multi-dimensional time series data
- reflect distributions of individual parameters
- include interactions/associations between parameters over time

- allow training models on synthetic data that will perform well on real data

GANs

Generative Adversarial Nets

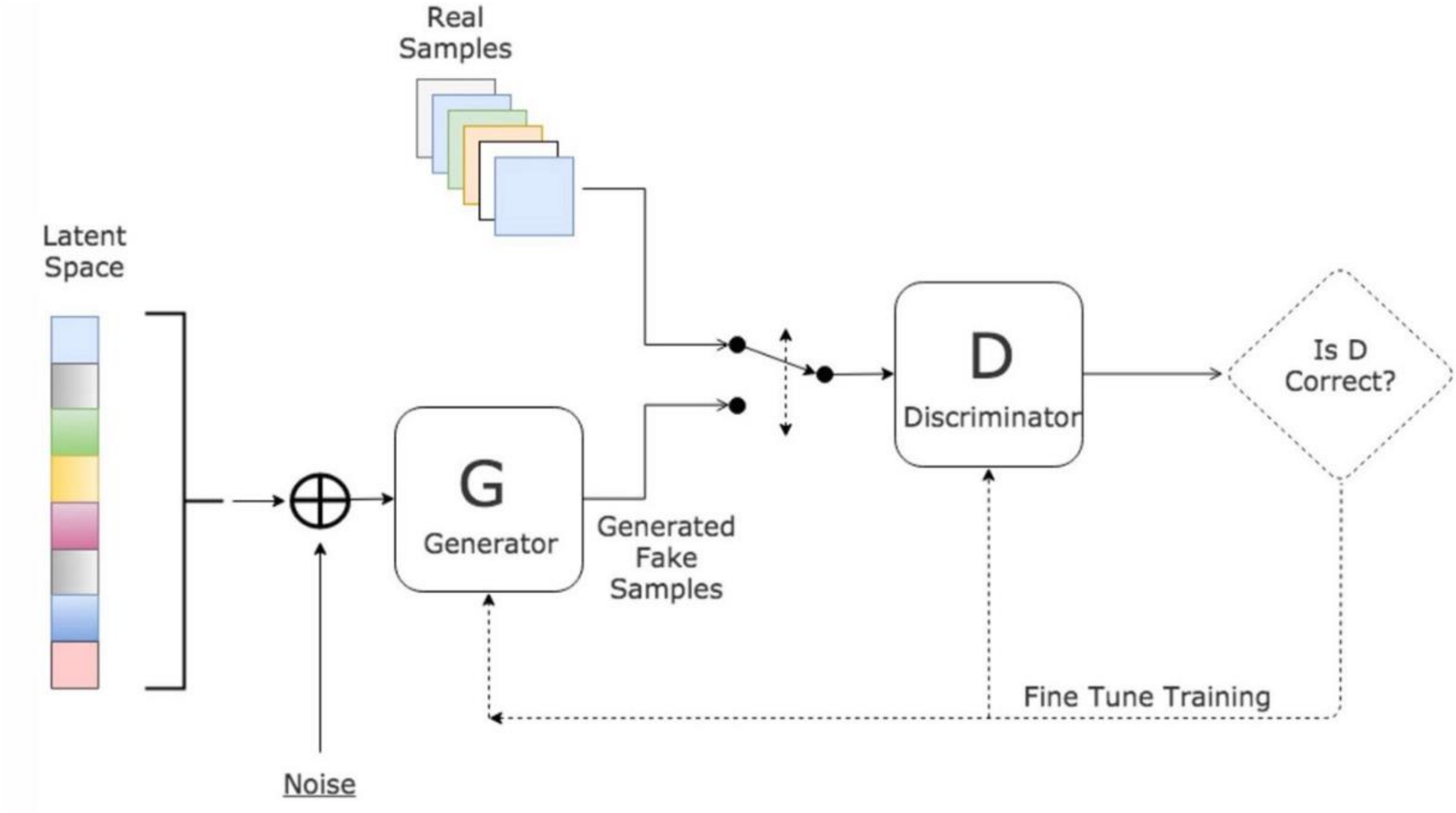
Ian J. Goodfellow, Jean Pouget-Abadie,* Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair,† Aaron Courville, Yoshua Bengio‡
Département d'informatique et de recherche opérationnelle
Université de Montréal
Montréal, QC H3C 3J7

Abstract

We propose a new framework for estimating generative models via an adversarial process, in which we simultaneously train two models: a generative model G that captures the data distribution, and a discriminative model D that estimates the probability that a sample came from the training data rather than G. The training procedure for G is to maximize the probability of D making a mistake. This framework corresponds to a minimax two-player game. In the space of arbitrary functions G and D, a unique solution exists, with G recovering the training data distribution and D equal to $\frac{1}{2}$ everywhere. In the case where G and D are defined by multilayer perceptrons, the entire system can be trained with backpropagation. There is no need for any Markov chains or unrolled approximate inference networks during either training or generation of samples. Experiments demonstrate the potential of the framework through qualitative and quantitative evaluation of the generated samples.

arXiv:1406.2661

structure of the given data, without specifying a target value. Generative models learn the intrinsic distribution function of the input data ***p(x)*** (or ***p(x,y)*** if there are multiple targets/classes in the dataset), allowing them to generate both synthetic inputs ***x'*** and outputs/targets ***y'***, typically given some hidden parameters.


Real
Samples
Latent
Space
G
Generator
Generated
Fake
Samples
D
Discriminator
Is D
Correct?
Noise
Fine Tune Training

Differentially Private Generative Adversarial Network

Liyang Xie[1], Kaixiang Lin[1], Shu Wang[2], Fei Wang[3], Jiayu Zhou[1]
[1]Computer Science and Engineering, Michigan State University
[2]Department of Computer Science, Rutgers University
[3]Department of Healthcare Policy and Research, Weill Cornell Medical School
{xieliyan,linkaixi}@msu.edu,sw498@cs.rutgers.edu,few2001@med.cornell.edu,jiayuz@msu.edu

ABSTRACT

Generative Adversarial Network (GAN) and its varian
cently attracted intensive research interests due to th
theoretical foundation and excellent empirical performa
erative models. These tools provide a promising direc
studies where data availability is limited. One comm
GANs is that the density of the learned generative d
could concentrate on the training data points, meanin
can easily *remember* training samples due to the high r
plexity of deep networks. This becomes a major con
GANs are applied to private or sensitive data such as pa
ical records, and the concentration of distribution ma
critical patient information. To address this issue, in thi
propose a differentially private GAN (DPGAN) model, in
achieve differential privacy in GANs by adding carefull
noise to gradients during the learning procedure. We p
orous proof for the privacy guarantee, as well as comp
empirical evidence to support our analysis, where we d
that our method can generate high quality data points a
able privacy level.

CCS CONCEPTS

• **Computing methodologies** → **Neural networks**; •
systems organization → **Neural networks**; • **Securit**
vacy → *Privacy-preserving protocols*;

KEYWORDS

Deep Learning; Differential Privacy; Generative model

arXiv:1802.06739

REAL-VALUED (MEDICAL) TIME SERIES GENERATION WITH RECURRENT CONDITIONAL GANS

Stephanie L. Hyland*
ETH Zurich, Switzerland
Tri-Institutional Training Program in Computational Biology and Medicine, Weill C
stephanie.hyland@inf.ethz.ch

Cristóbal Esteban*
ETH Zurich, Switzerland
cristobal.esteban@inf.ethz.ch

Gunnar Rätsch
ETH Zurich, Switzerland
raetsch@inf.ethz.c

ABSTRACT

Generative Adversarial Networks (GANs) have shown remarkable success
framework for training models to produce realistic-looking data. In this work
propose a Recurrent GAN (RGAN) and Recurrent Conditional GAN (RCGA
produce realistic *real-valued multi-dimensional time series*, with an emphas
their application to medical data. RGANs make use of recurrent neural netw
(RNNs) in the generator and the discriminator. In the case of RCGANs, bo
these RNNs are conditioned on auxiliary information. We demonstrate our m
in a set of toy datasets, where we show visually and quantitatively (using sa
likelihood and maximum mean discrepancy) that they can successfully gen
realistic time-series. We also describe novel evaluation methods for GANs, w
we generate a synthetic labelled training dataset, and evaluate on a *real te*
the performance of a model trained on the *synthetic data*, and vice-versa.
illustrate with these metrics that RCGANs can generate time-series data u
for supervised training, with only minor degradation in performance on *rea*
data. This is demonstrated on digit classification from 'serialised' MNIST ar
training an early warning system on a medical dataset of 17,000 patients fro
intensive care unit. We further discuss and analyse the privacy concerns that
arise when using RCGANs to generate realistic synthetic medical time series
and demonstrate results from differentially private training of the RCGAN.

arXiv:1706.02633

Privacy-preserving generative deep neural networks support clinical data sharing

Authors: Brett K. Beaulieu-Jones[1], Zhiwei Steven Wu[2], Chris Williams[3], Casey S. Greene[3]*

[1]Genomics and Computational Biology Graduate Group, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, USA.
[2]Computer and Information Science, School of Engineering and Applied Sciences, University of Pennsylvania, Philadelphia, Pennsylvania, USA.
[3]Department of Systems Pharmacology and Translational Therapeutics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, USA.

*To whom correspondence should be addressed: csgreene@upenn.edu

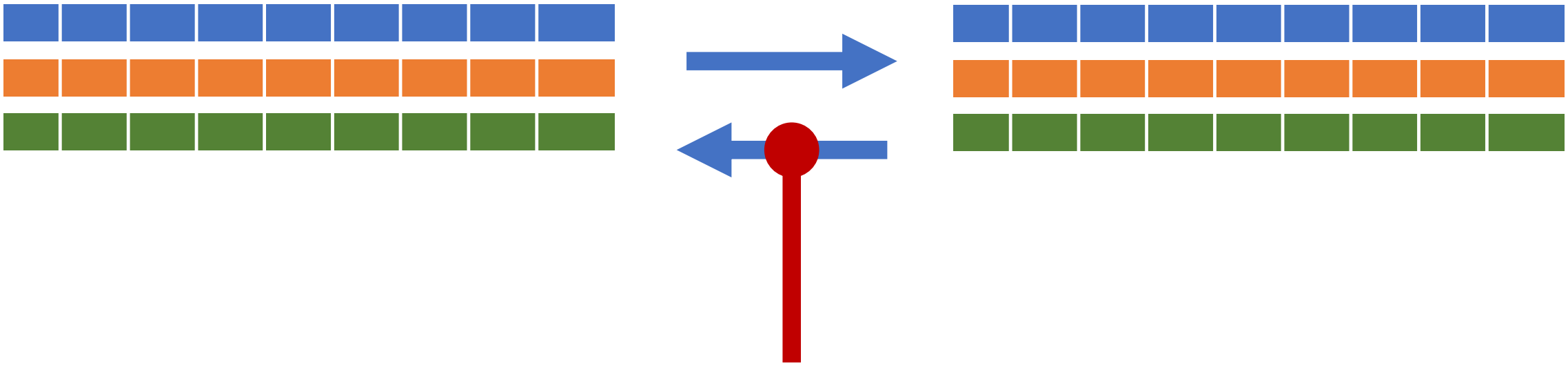
One Sentence Summary: Deep neural networks can generate shareable biomedical data to allow reanalysis while preserving the privacy of study participants.

Abstract: Though it is widely recognized that data sharing enables faster scientific progress, the sensible need to protect participant privacy hampers this practice in medicine. We train deep neural networks that generate synthetic subjects closely resembling study participants. Using the SPRINT trial as an example, we show that machine-learning models built from simulated participants generalize to the original dataset. We incorporate differential privacy, which offers strong guarantees on the likelihood that a subject could be identified as a member of the trial. Investigators who have compiled a dataset can use our method to provide a freely accessible public version that enables other scientists to perform discovery-oriented analyses. Generated data can be released alongside analytical code to enable fully reproducible workflows, even when privacy is a concern. By addressing data sharing challenges, deep neural networks can facilitate the rigorous and reproducible investigation of clinical datasets.

https://doi.org/10.1101/159756

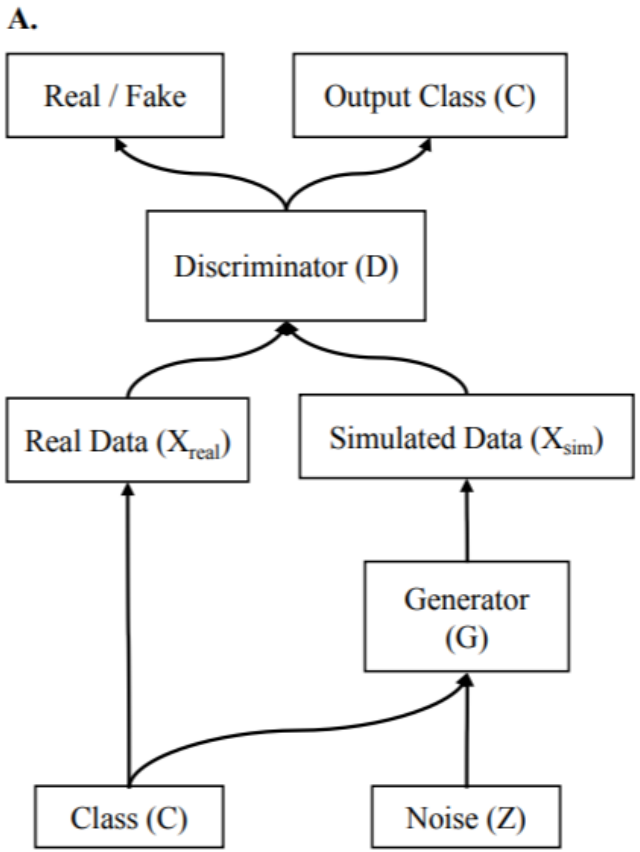
real time-series data

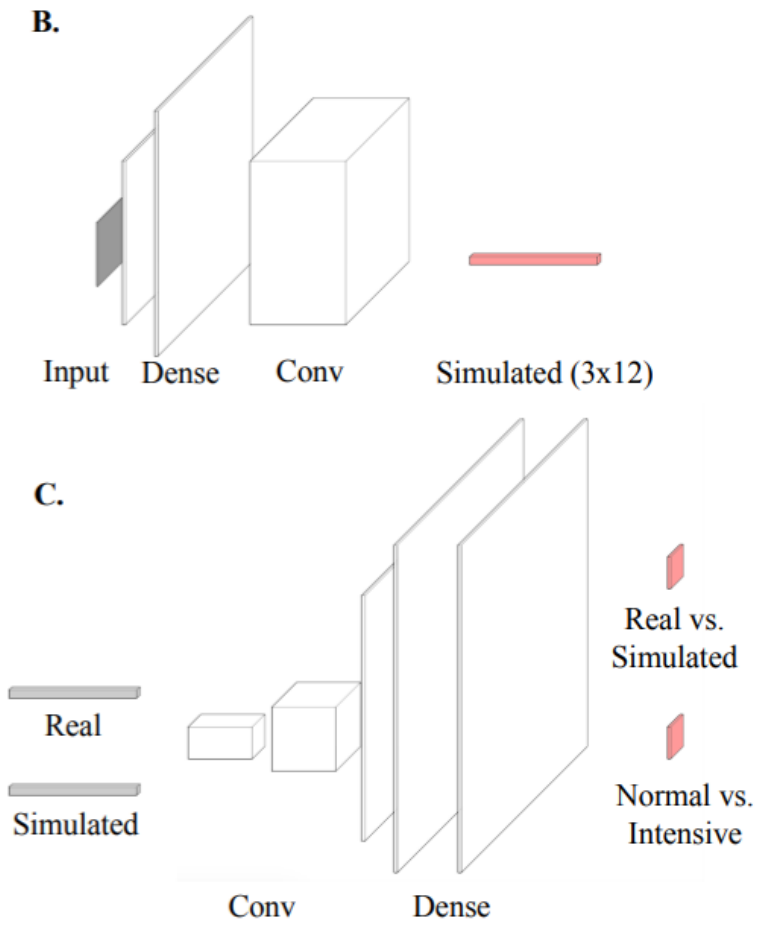
synthetic time-series data



differential privacy inhibits data leakage

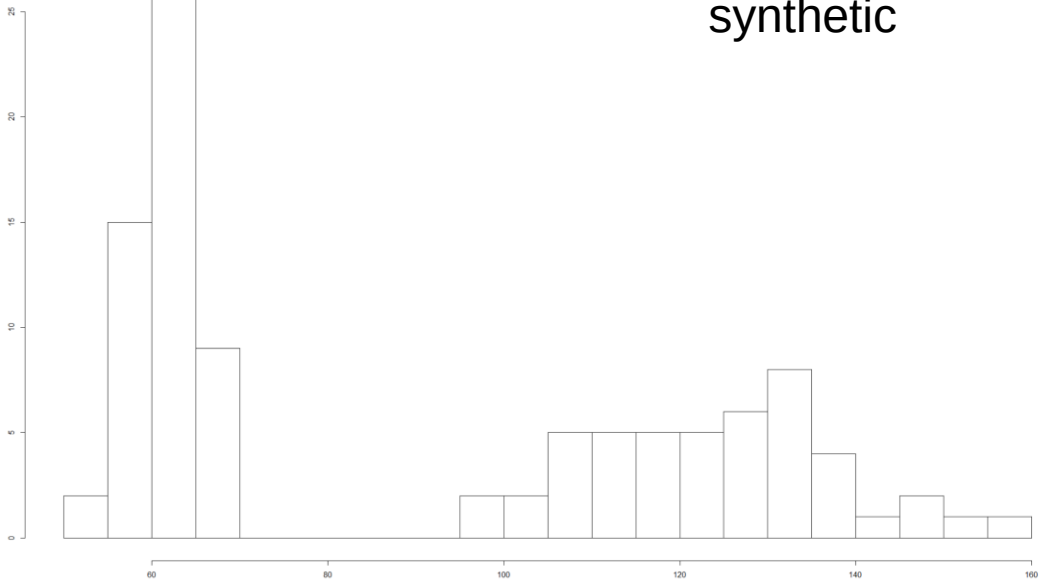
- quality vs security tradeoff
- privacy budgets
- computational limits



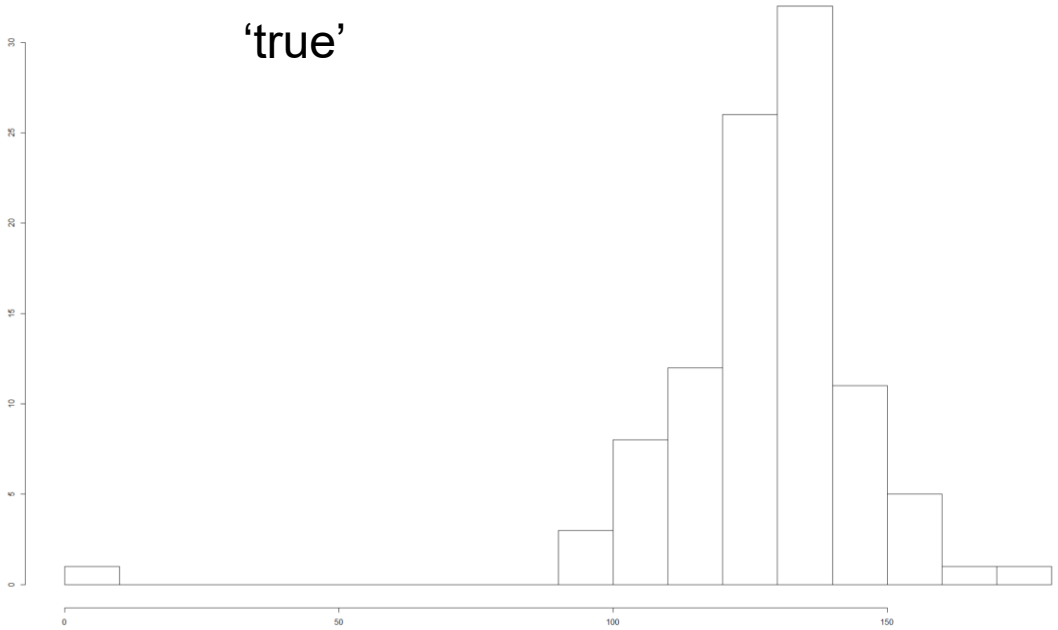


beaulieu-jones et al 2017. doi: https://doi.org/10.1101/159756

synthetic

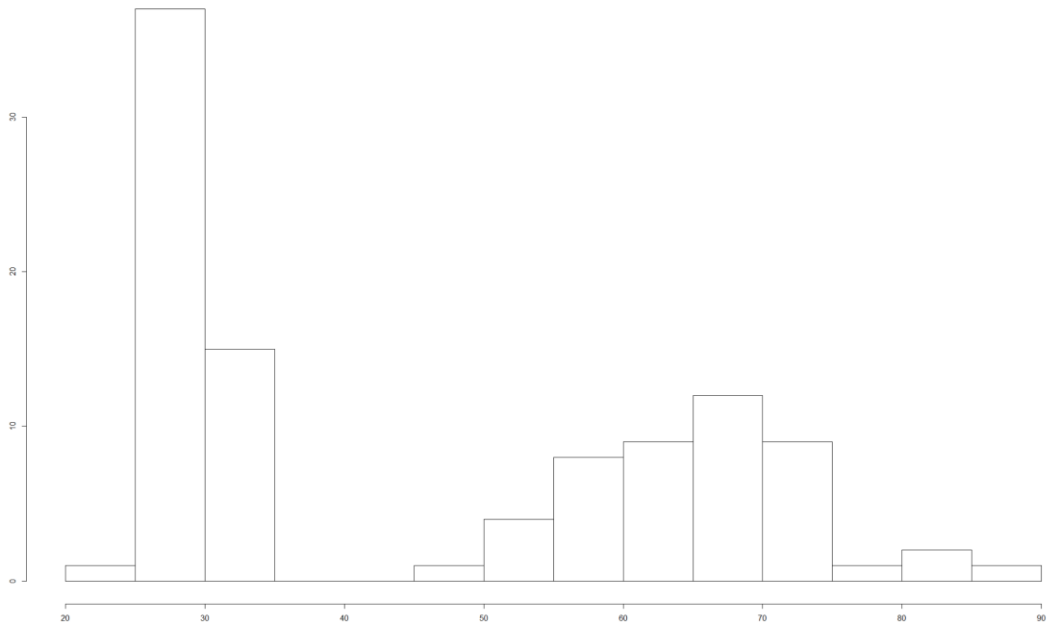


‘true’

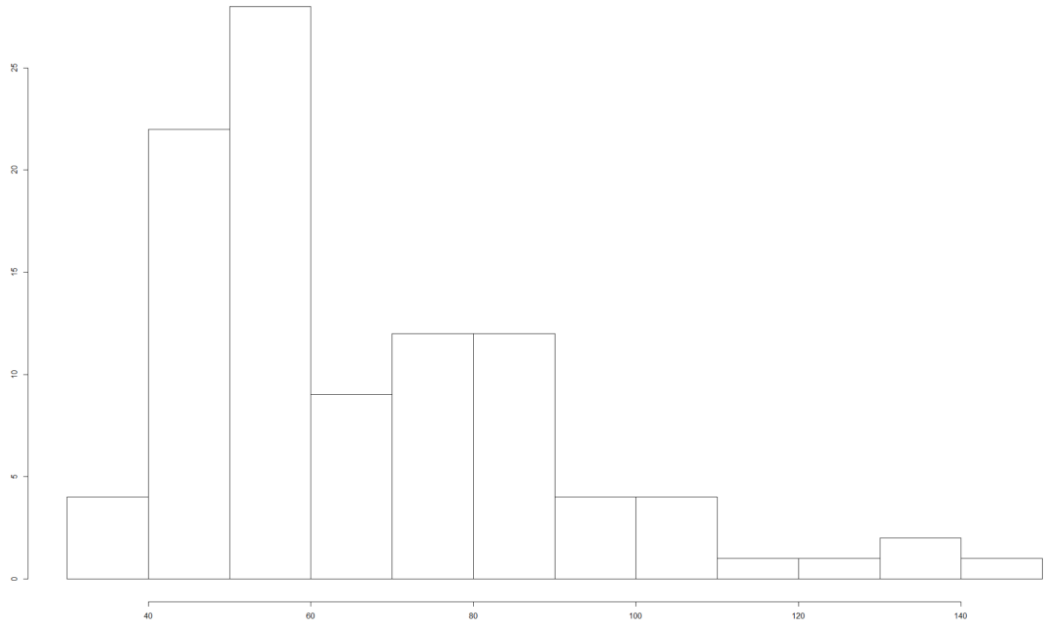


GAN vs true generated systolic BP distributions – 100 epochs with differential privacy
single time point, 100 ‘IDs’

synthetic



‘true’



GAN vs true generated HbA1c distributions – 100 epochs with differential privacy
single time point, 100 ‘IDs’

potential uses of GANs generated synthetic data

- balance classes (upsampling)
- with dp implemented – allow data sharing

- with further development can be used as classifier/regressors – may provide a general solution to all problems discussed

glucose.ai group, NHS GGC
Greg Jones
Debbie Morrison
Sean Harbison

Stratified Medicine MSc students
Ruth Muir
Alastair Irvine
Halah al Saadi

myDiabetesIQ
Debbie Wake
Scott Cunningham
Nicky Conway
Sam Philip
Felix Agakov

thinkingAI group, University of Birmingham
Krish Nirantharakumar
Krishna Gokhale
Tom Taverner
Peter Tino