personalising risk prediction in diabetes using machine learning
bridging the gap between RCT and RCD

Chris Sainsbury
Queen Elizabeth University Hospital, Glasgow
NHS Research Scotland (NRS) Senior Fellow
Hon Senior Research Fellow, University of Birmingham

@csainsbury
http://glucose.ai

NHS Greater Glasgow and Clyde
NHS SCOTLAND NHS RESEARCH SCOTLAND
UNIVERSITY OF BIRMINGHAM
MyWay digital health
glucose.ai group
Innovate UK


UNIVERSITY OF BIRMINGHAM
Teaching excellence
Alumni
Work here
Events
Study
Research
International
Business
Home > Research > Institute of Applied Health Research
Research Spotlights
Our Researchers
Research Areas
Institutes
News
Institute of Applied Health Research

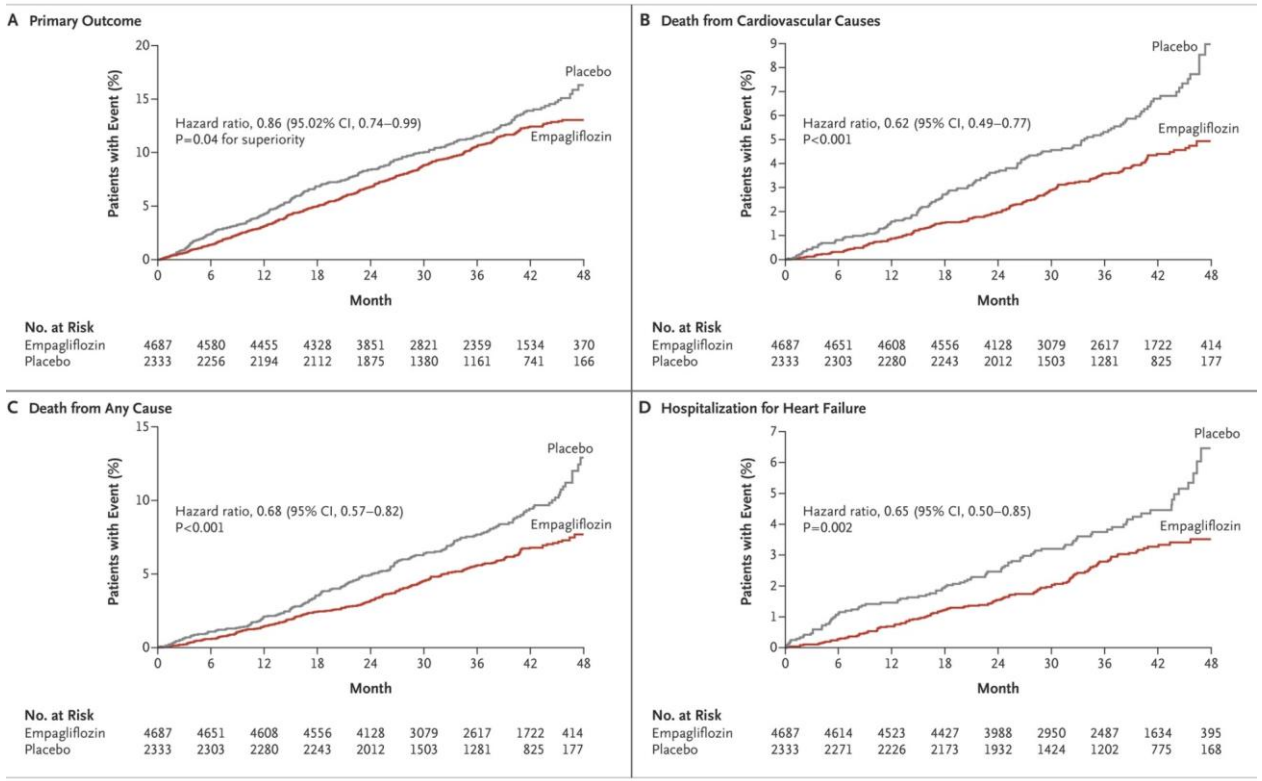
North Sea
Dundee
Edinburgh
Glasgow
United Kingdom
Newcastle upon Tyne
Isle of Man
Leeds
Manchester
Liverpool
Dublin
Galway
Ireland
Limerick
Cork
WALES
Oxford
Bristol
Cardiff
Brighton
Southampton
Brussels
Lille
Belgium

Machine learning

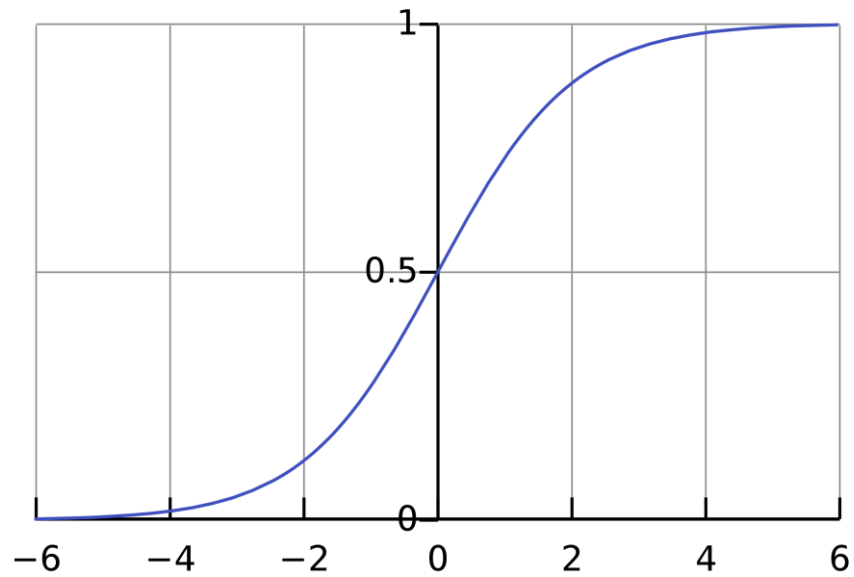
From Wikipedia, the free encyclopedia

Machine learning (**ML**) is the scientific study of algorithms and statistical models that computer systems use to effectively perform a specific task without using explicit instructions, relying on patterns and inference instead. It is seen as a subset of artificial intelligence.

- Predict outcomes for individuals vs populations
- Access information from complete dataset (TS / text etc)



population level outcomes



Logit as basis for many current clinical risk scoring systems

Admission Glucose Number (AGN): A Point of Admission Score Associated With Inpatient Glucose Variability, Hypoglycemia, and Mortality

Journal of Diabetes Science and Technology
2019, Vol. 13(2) 213–220

DOI: 10.1177/1932296818800722
journals.sagepub.com/home/dst
SAGE

Jennifer McKechnie, BSc, MBChB[1], Rahat Maitland, MD, FRCP[1], Christopher A. R. Sainsbury, BSc, MD, FRCP, FFCI[1], and Gregory C. Jones, MBChB, FRCP[1]

Received: 13 October 2017 | Revised: 6 December 2017 | Accepted: 17 December 2017
DOI: 10.1111/dom.13193
WILEY

ORIGINAL ARTICLE

Visit-to-visit HbA1c variability and systolic blood pressure (SBP) variability are significantly and additively associated with mortality in individuals with type 1 diabetes: An observational study

Stuart S. Wightman | Christopher A. R. Sainsbury MD | Gregory C. Jones MBChB

DIABETICMedicine
DOI: 10.1111/dme.13621

Short Report: Educational and Psychological Aspects

Structured education using Dose Adjustment for Normal Eating (DAFNE) reduces long-term HbA_{1c} and HbA_{1c} variability

G. S. Walker[1], J. Y. Chen[1], H. Hopkinson[2], C. A. R. Sainsbury[1] and G. C. Jones[1]

[1]Diabetes Centre, Gartnavel General Hospital and [2]Diabetes Centre, New Victoria Hospital, Glasgow, UK

Accepted 9 March 2018

Hypoglycemia and Clinical Outcomes in Hospitalized Patients With Diabetes: Does Association With Adverse Outcomes Remain When Number of Glucose Tests Performed Is Accounted For?

Journal of Diabetes Science and Technology
2017, Vol. 11(4) 720–723

DOI: 10.1177/1932296816688012
journals.sagepub.com/home/dst
SAGE

Gregory C. Jones, MB ChB[1], Joseph G. Timmons, MB ChB[1], Scott G. Cunningham, PhD[2], Stephen J. Cleland, MD[1], and Christopher A. R. Sainsbury, MD[1]

? more ambitious use of data resources available

SCI-Diabetes

>99% population coverage

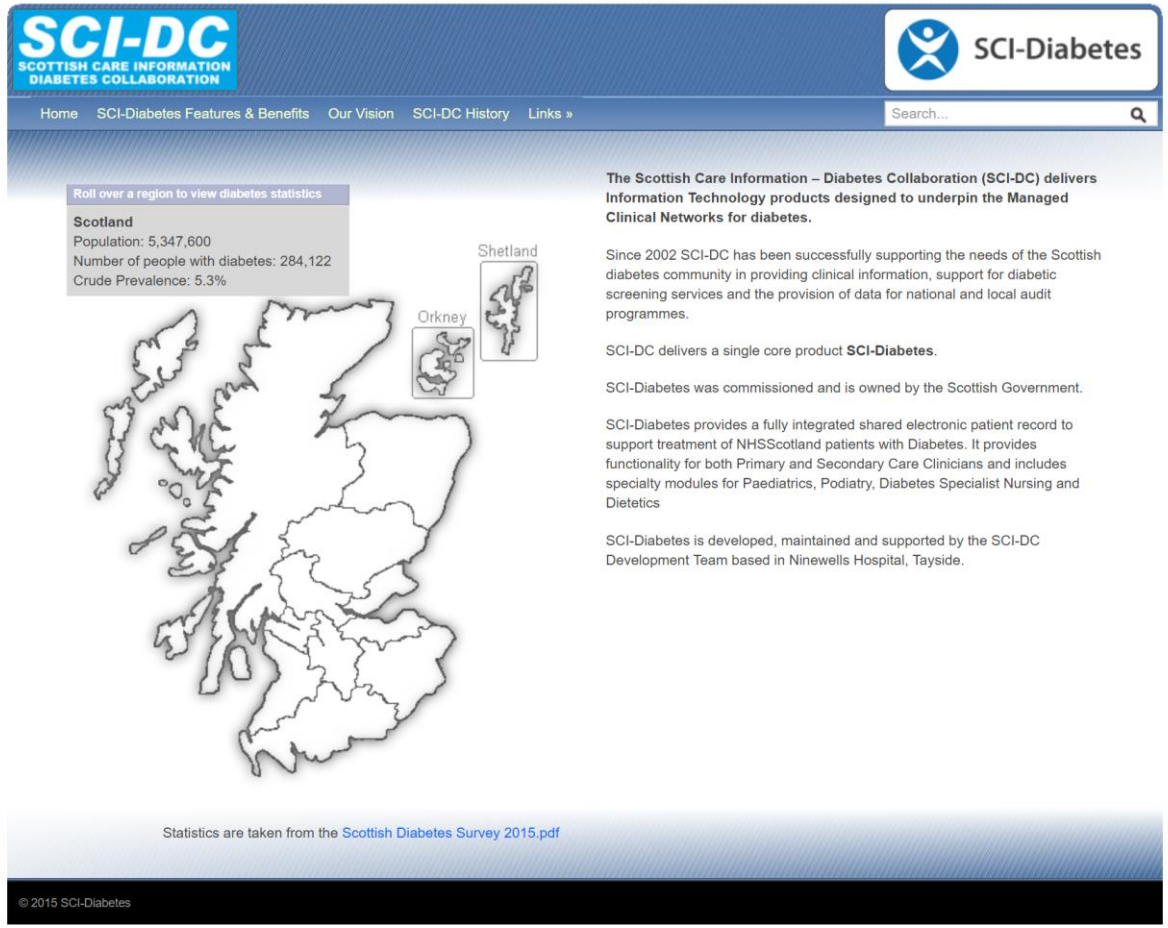
~120 000 IDs with DM NHS GGC

comprehensive dataset

primary/secondary care results:

phenotyping information

HbA1c, BMI, BP, bloods, prescriptions



data context

i. Large amounts of routinely collected data (RCD) are **unused** in clinical decision making

ii. Clinical decisions are largely made on the basis of summary information (averages / snapshots)

iii. Clinical guidelines are based on randomised controlled trial (RCT) data

iv. RCTs not representative of the vast majority of the real world population (comorbid, elderly individuals excluded)

v. RCD (retrospective cohort studies) analyses not highly regarded due to potential hidden biases etc

we think in RCT-space, but work in RCD-space

overall approach

i. Tackle the problem of summary data use by constructing analyses that use time series information

ii. Extend analyses into additional data types (text)

iii. Calibrate RCD cohort investigations to RCT trial data – understand associations and biases

iv. Explore the effect size of interventions in populations that reflect the real world (comorbid, elderly etc)

predict the response (effect size within a specified domain) to an **arbitrary intervention** in an **arbitrary population** (or individual)

report effect size in both RCD and RCT-space

- predict optimum therapy choice for an individual / population (decision support)
- predict at-risk populations for adverse events (decision support)
- investigate potential for indication expansion for existing therapies (trial design)
- Investigate sources and size of hidden biases

CASE 1

58 y male

Type 2 Diabetes - 8y duration

No other diagnoses

HbA1c 67

CASE 2

73 y male

Type 2 Diabetes - 3y duration

COPD

HbA1c 67

CASE 3

80 y female

Type 2 Diabetes - 20y duration

COPD, Psoriatic Arthritis, Ca Breast

HbA1c 67

Increasing complexity

More like real world

Less like evidence / guidelines

Harder clinical decision making

CASE 1

58 y male
Type 2 Diabetes - 8y duration
No other diagnoses
HbA1c 67

Case 1

clinical context simple
actual decision making remains difficult
guidelines (in DM at least) ambiguous / open
fine for specialists – not so much for generalists / primary care etc

theme 1 **drug response prediction in diabetes:**
(virtual n=1 drug trial)
myDiabetesIQ

related projects – DeepMind collaboration / similarOme

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

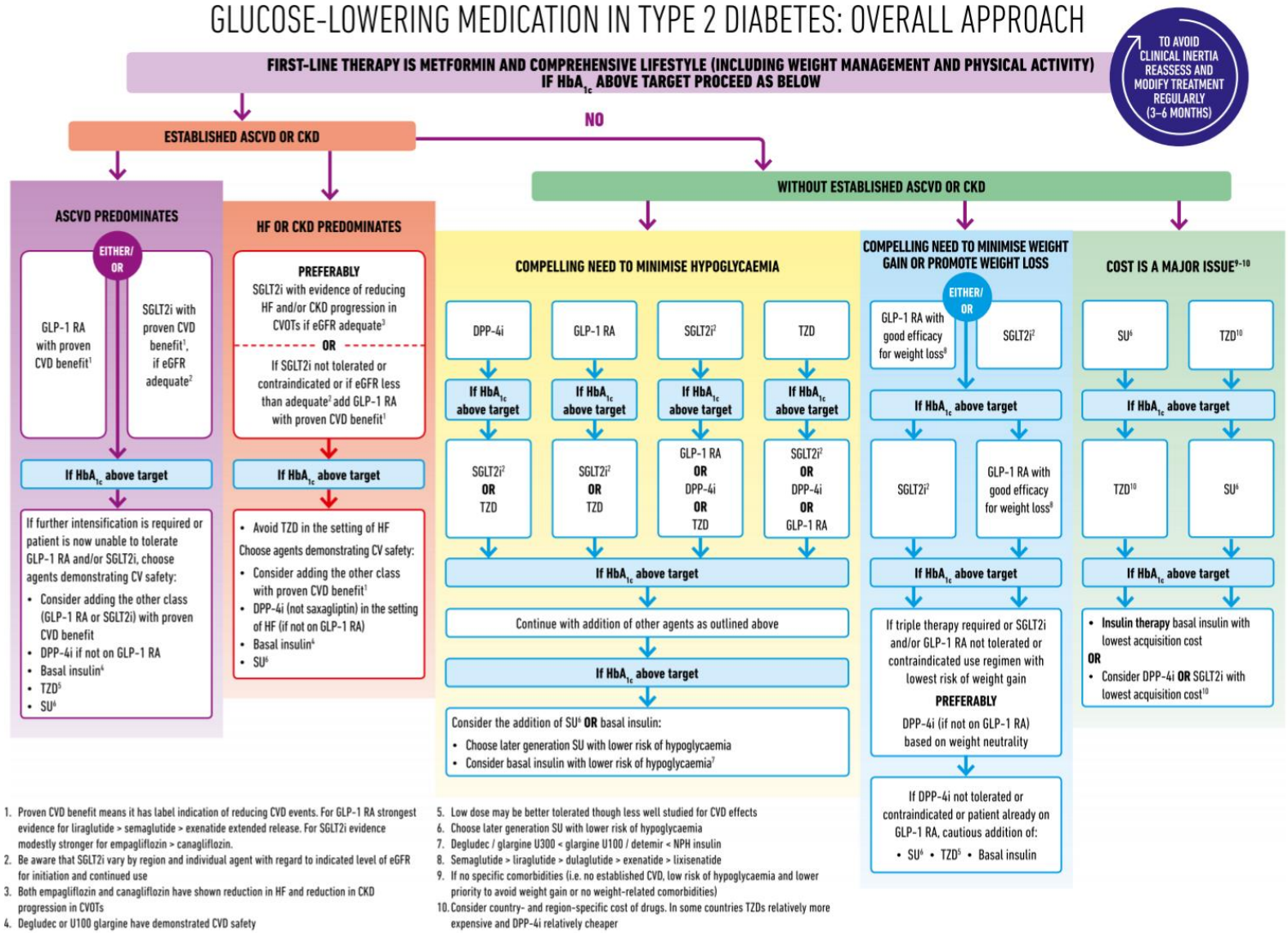


Fig. 2 Glucose-lowering medication in type 2 diabetes: overall approach

ADA / EASD, 2018

Davies, Melanie J., et al . 2018. Diabetologia 61 (12): 2461–98.

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

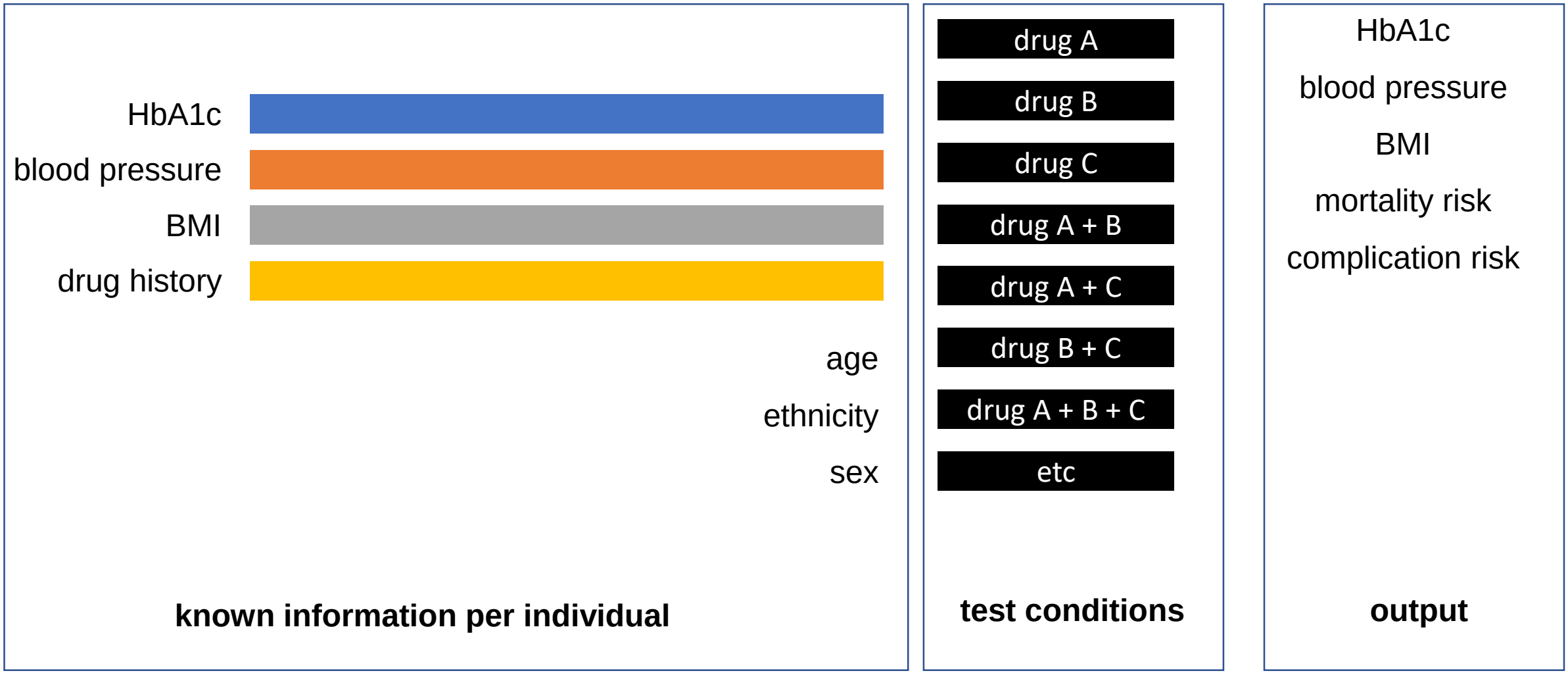
time series

- sex
- age
- ethnicity

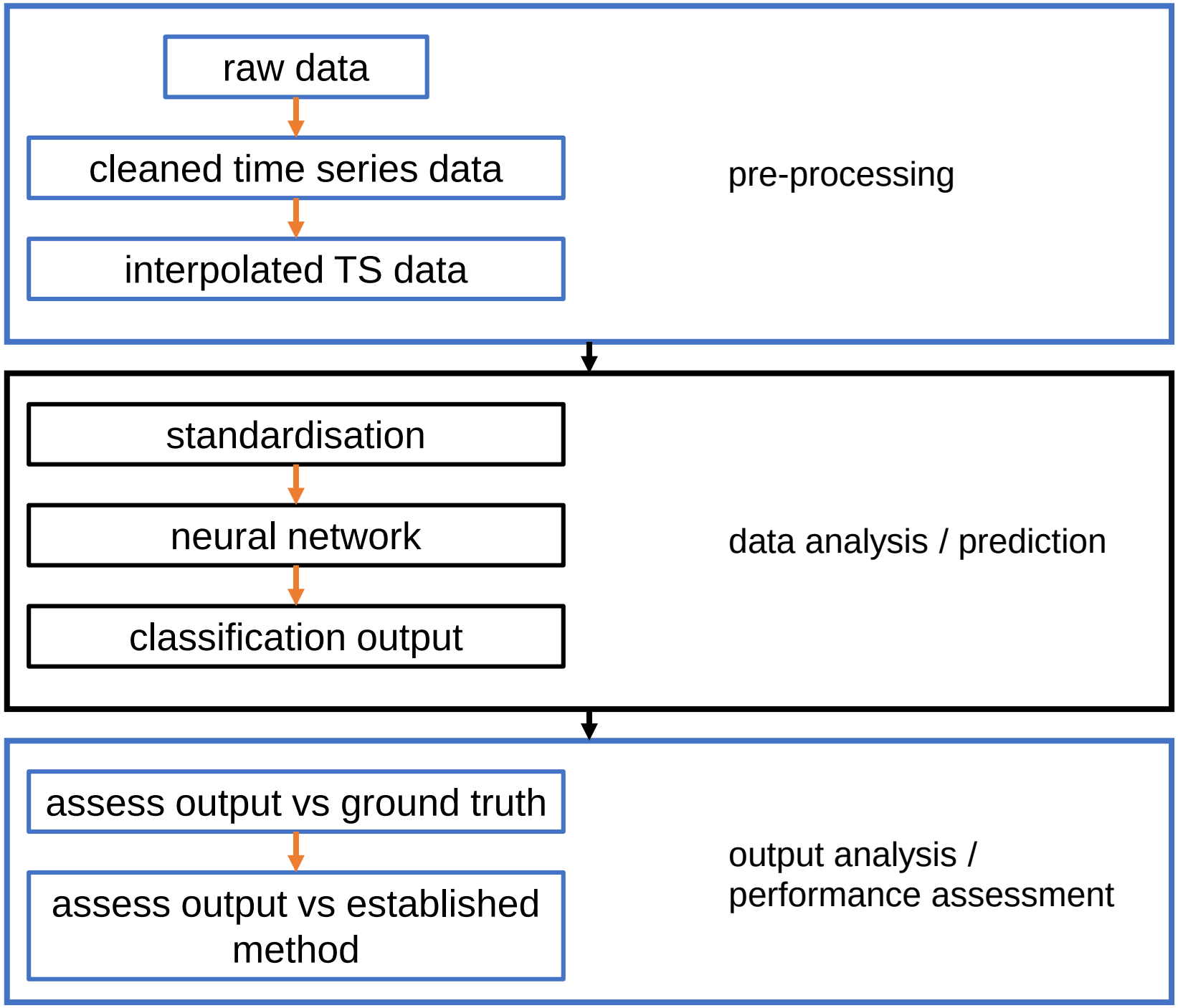
stable over time

Methods – general approach

predict

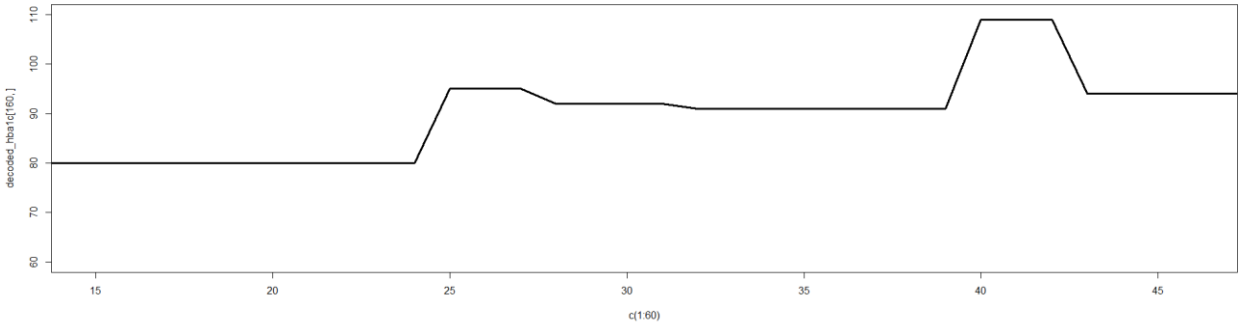


Workflow

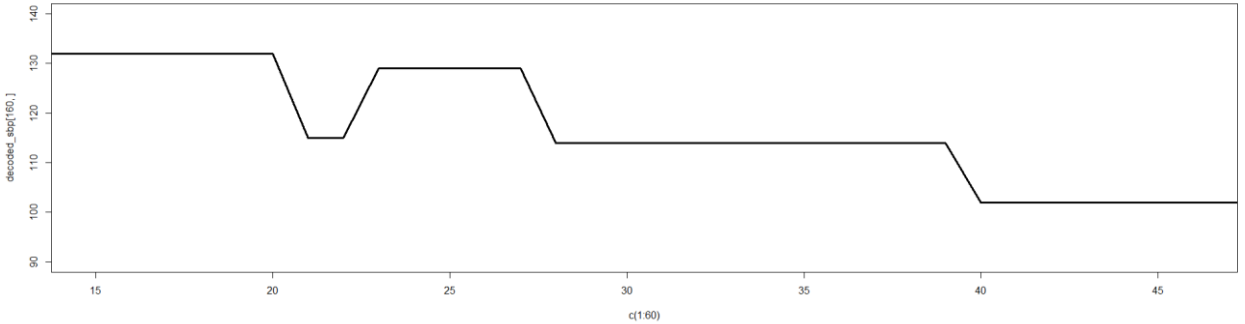

raw data
cleaned time series data
interpolated TS data
pre-processing
standardisation
neural network
classification output
data analysis / prediction
assess output vs ground truth
assess output vs established method
output analysis / performance assessment

managing time series data – numerical data

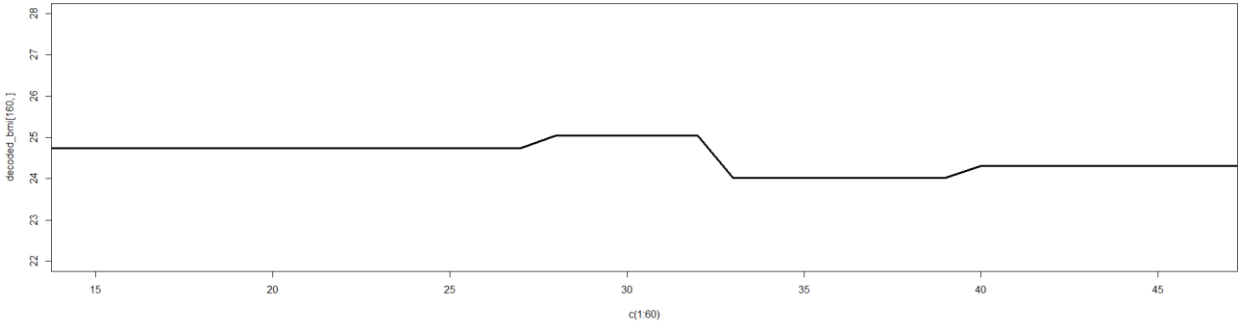
hba1c



sbp



bmi

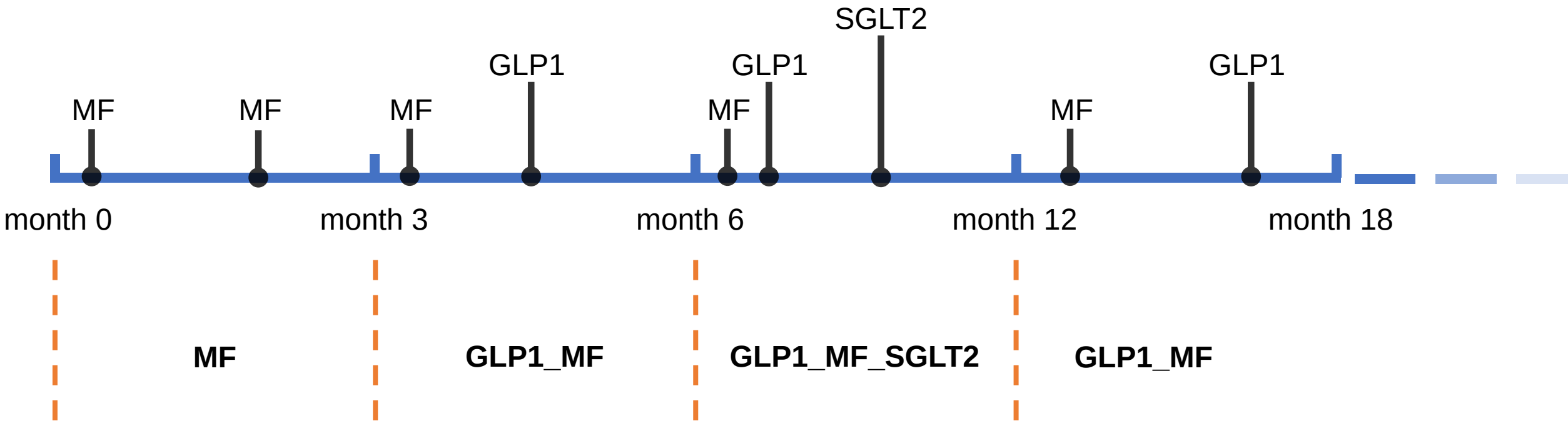


time

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

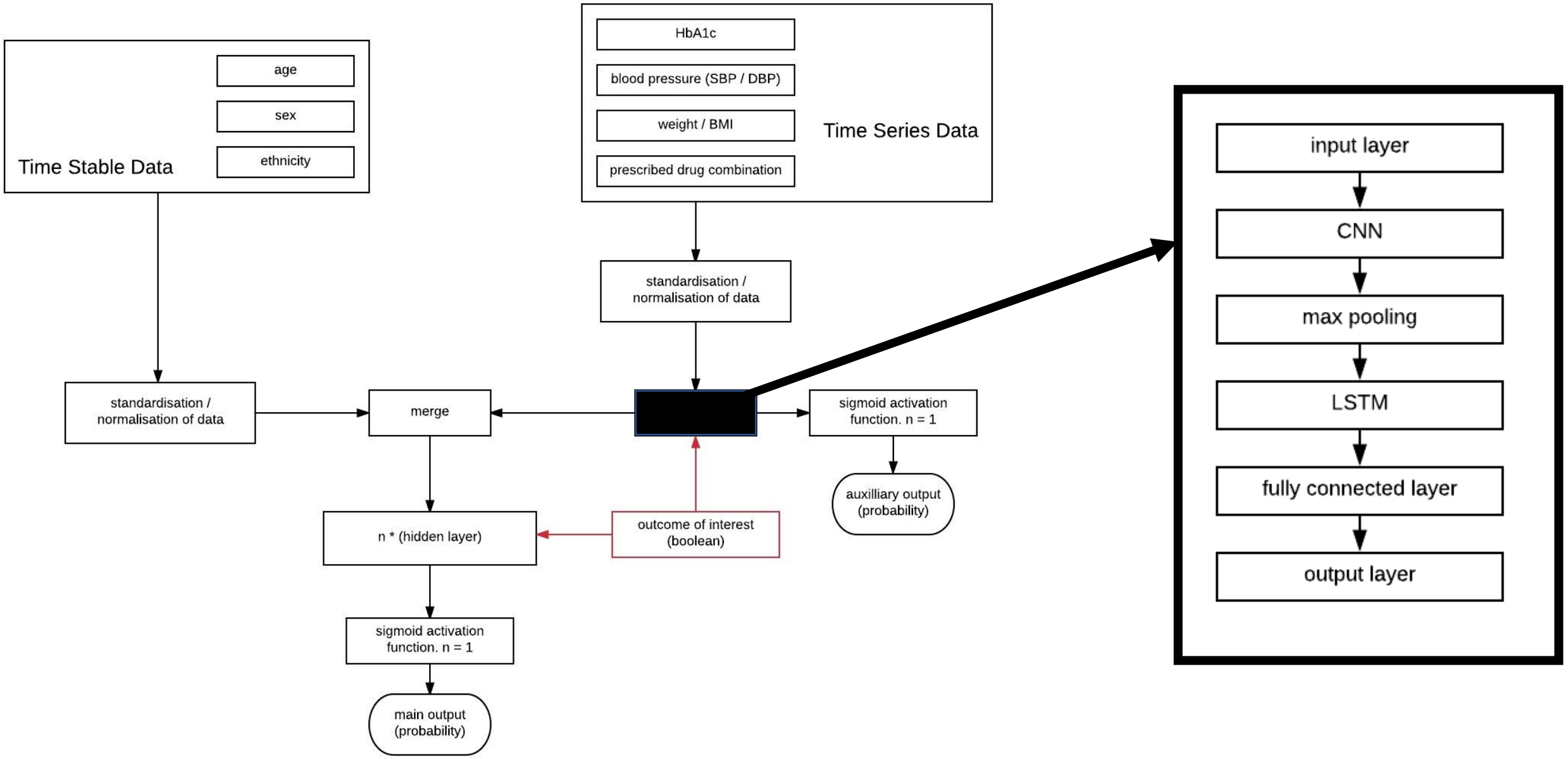
↓

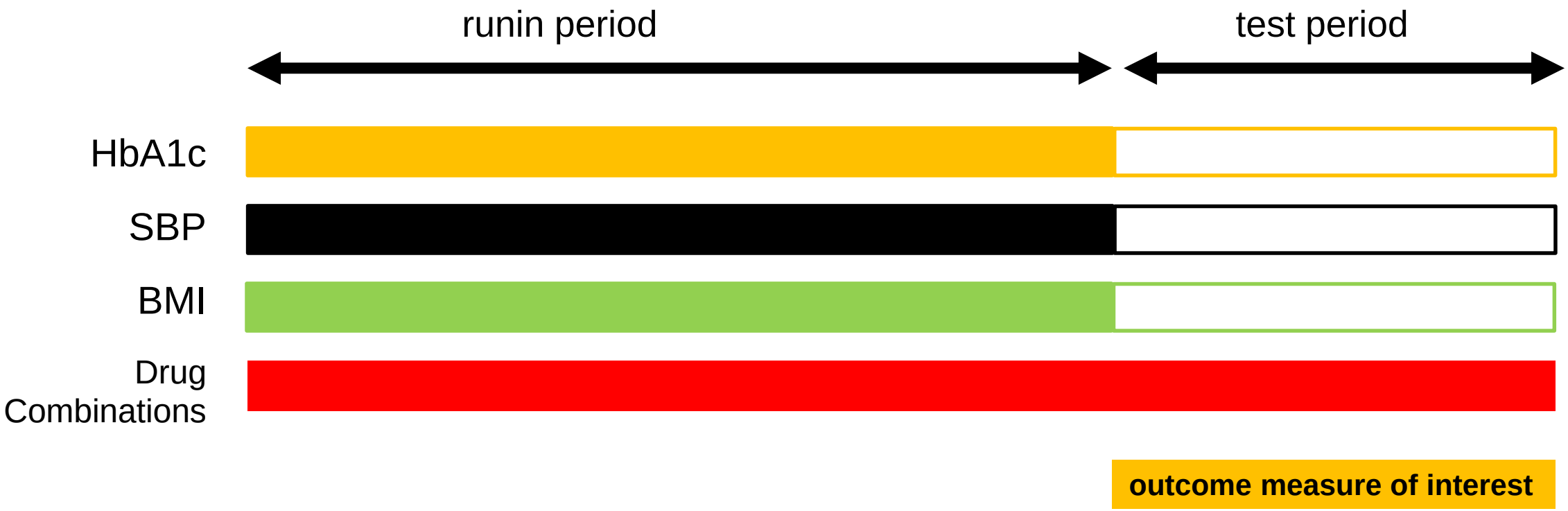
Embedding -> numerical vector

↓

input into RNN / LSTM

schematic of neural network classifier





eg

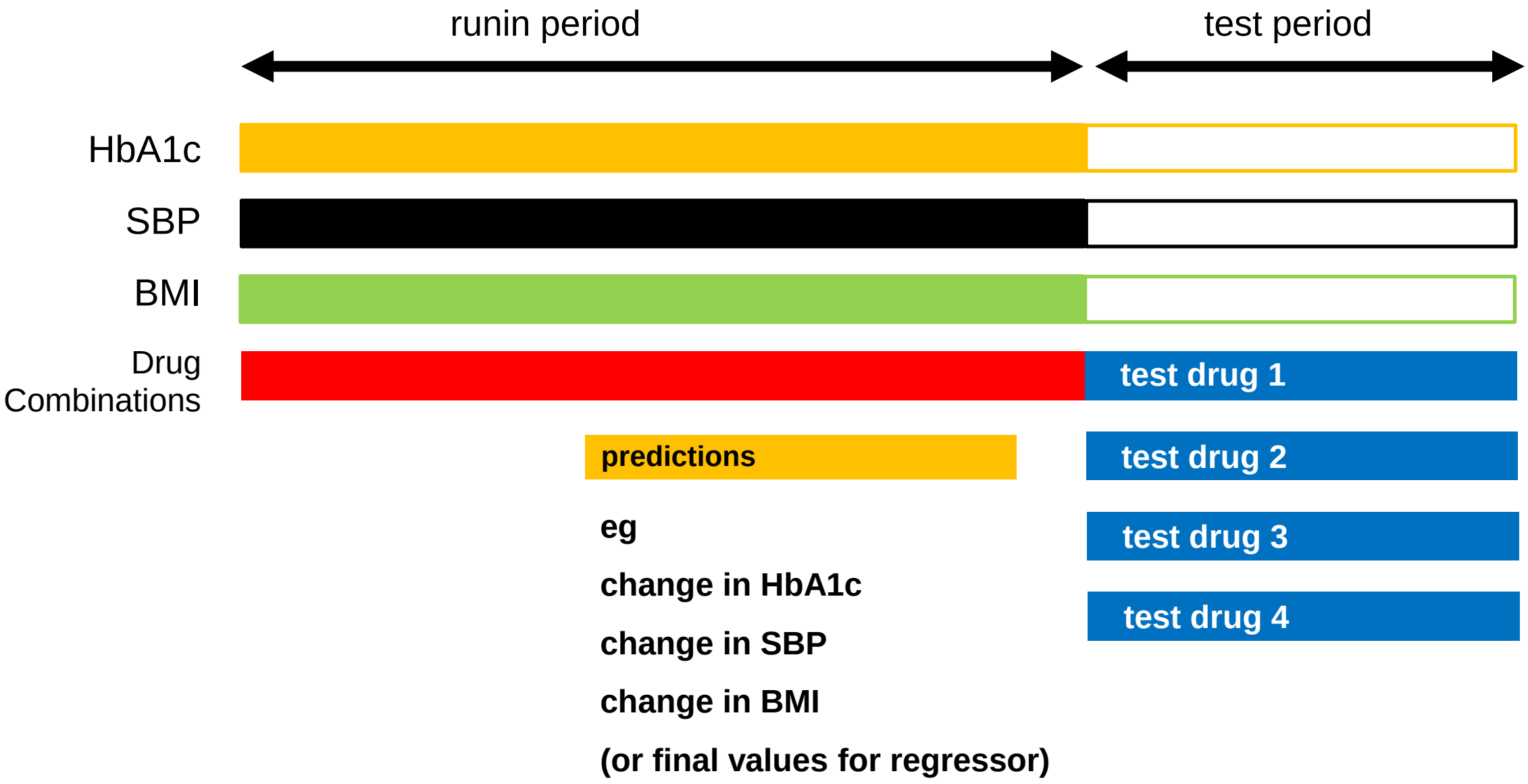
change in HbA1c

change in SBP

change in BMI

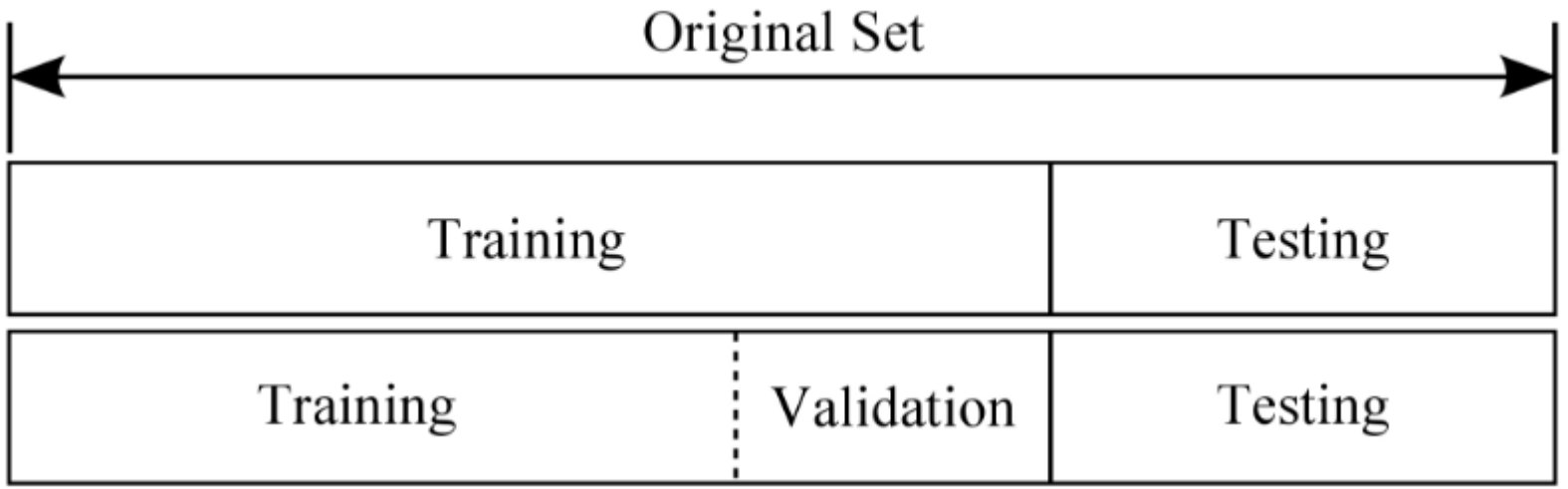
(or time-point standardised values for regressor version)

training data structure



using the model to predict response

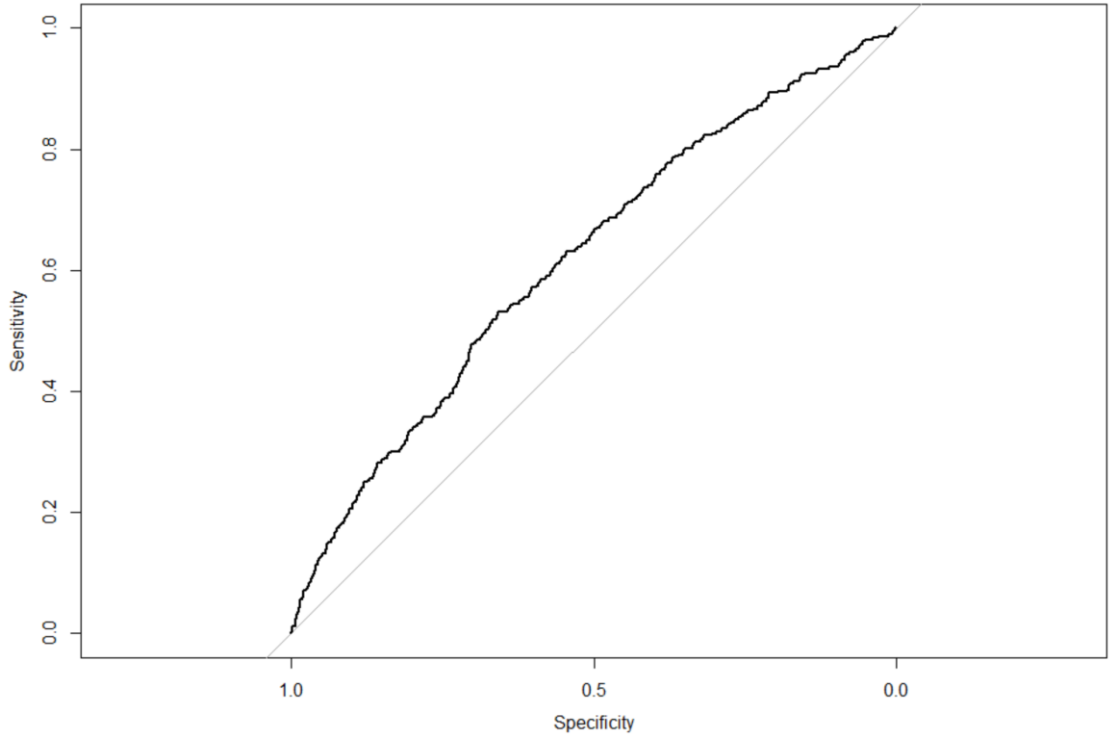
training, validation and withheld test sets



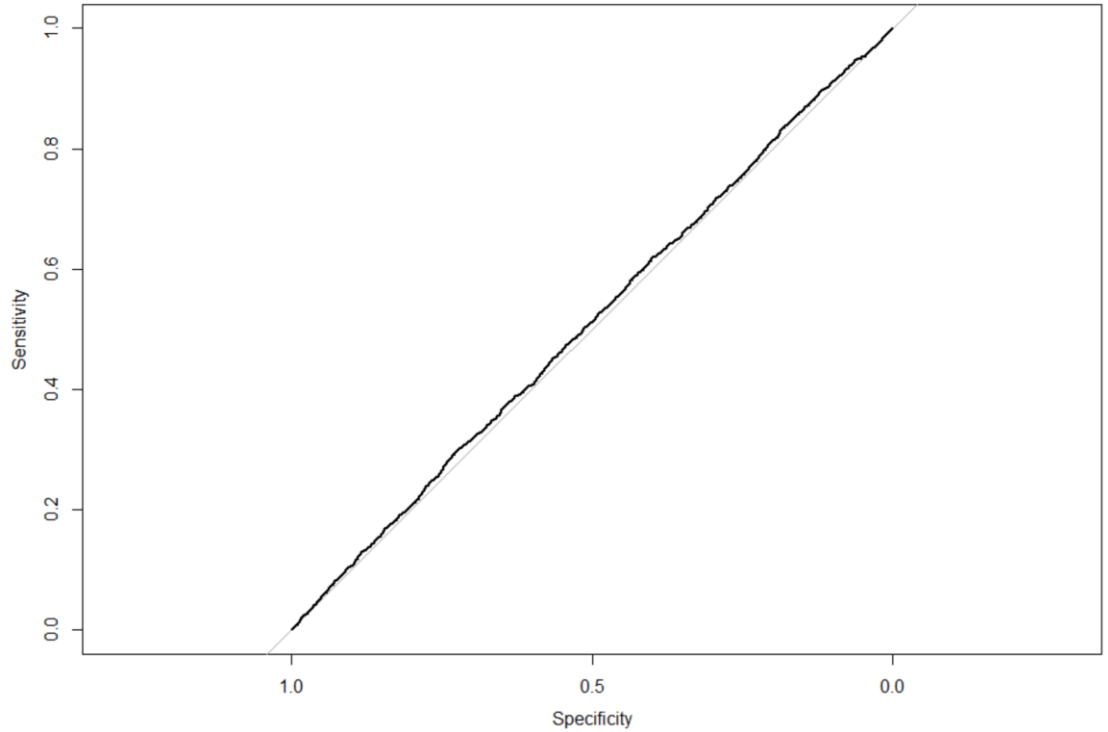
nn output vs logistic regression
HbA1c TS vs HbA1c median / CV

predicting mortality at 3y

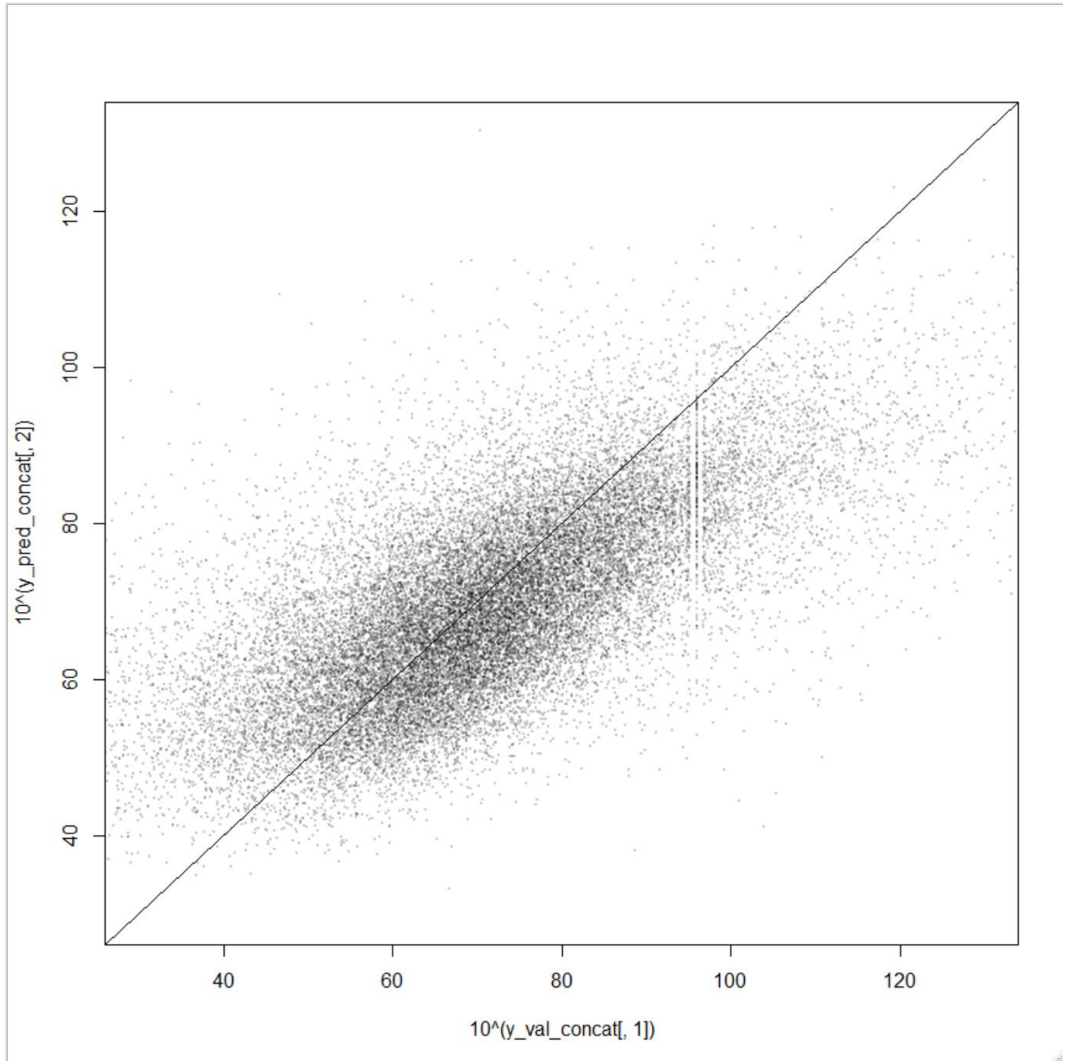
8 year runin analysis : outcome: all-cause mortality at 3y

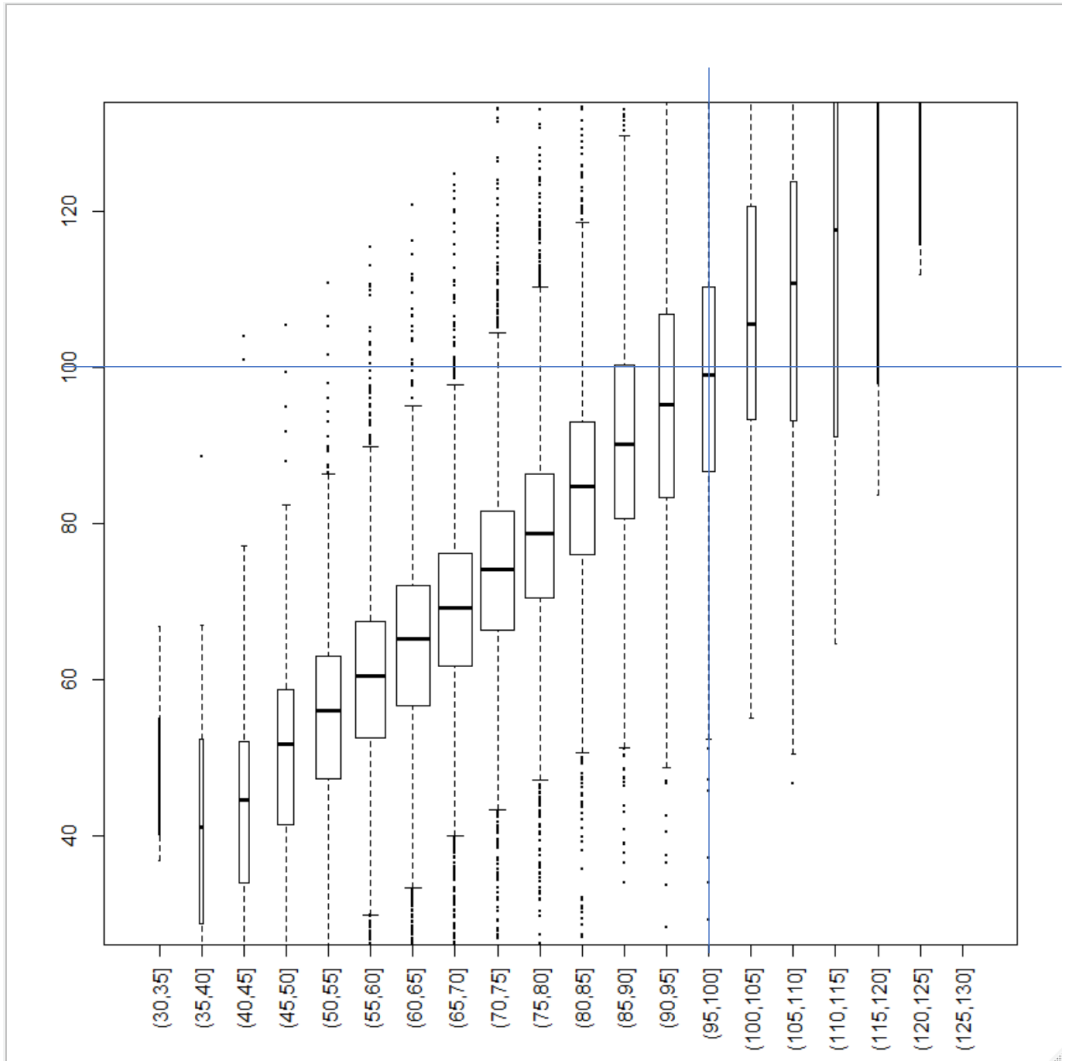


nn output
AUROC 0.6



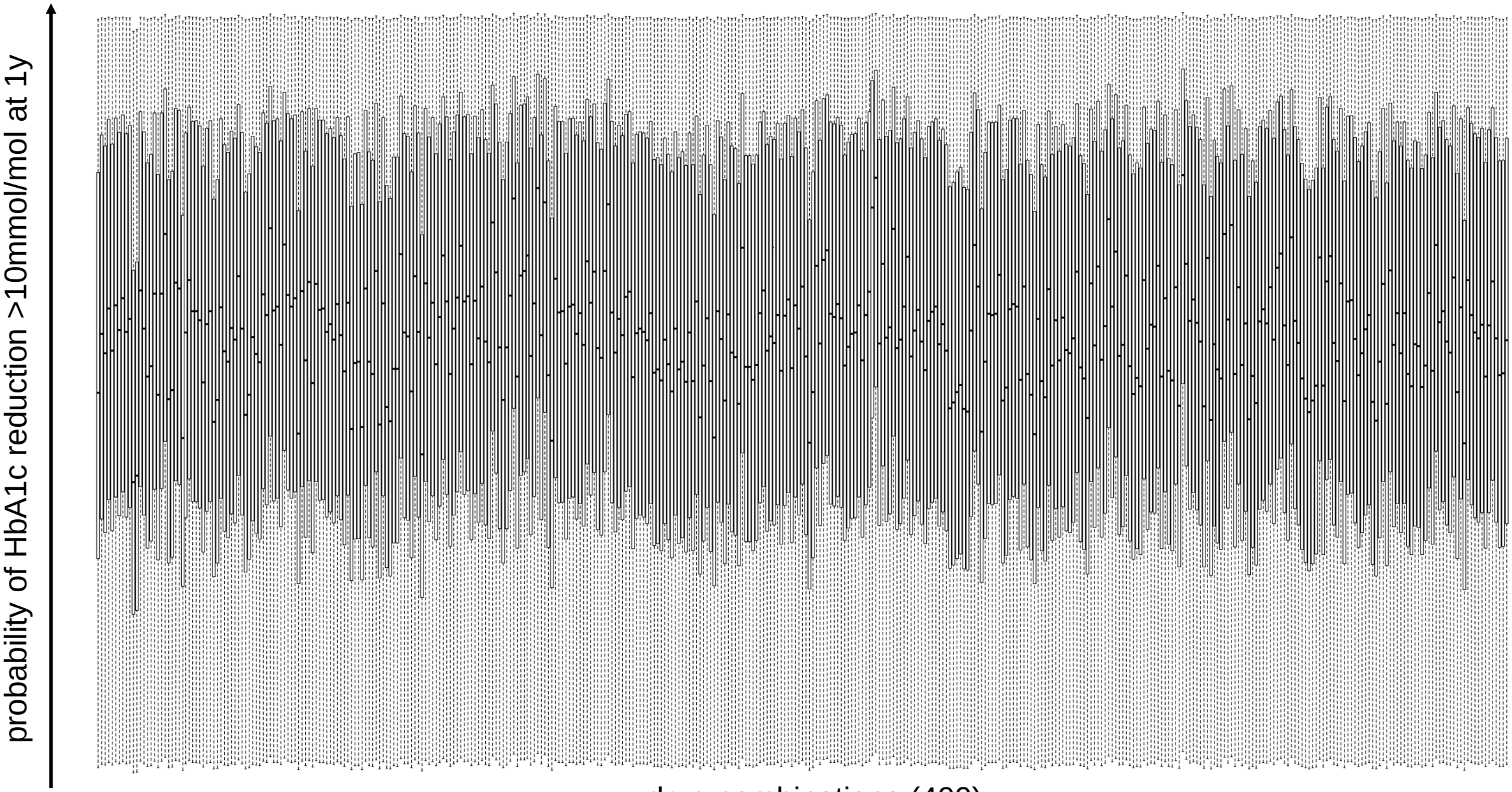
logistic regression
AUROC 0.51





T1 multistep n=26860
target = gradient of line

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



Insight from text data

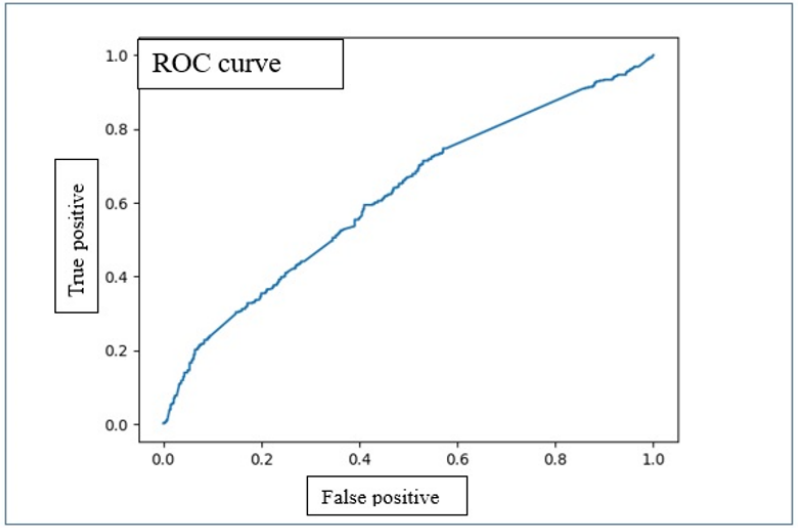
free-text entries

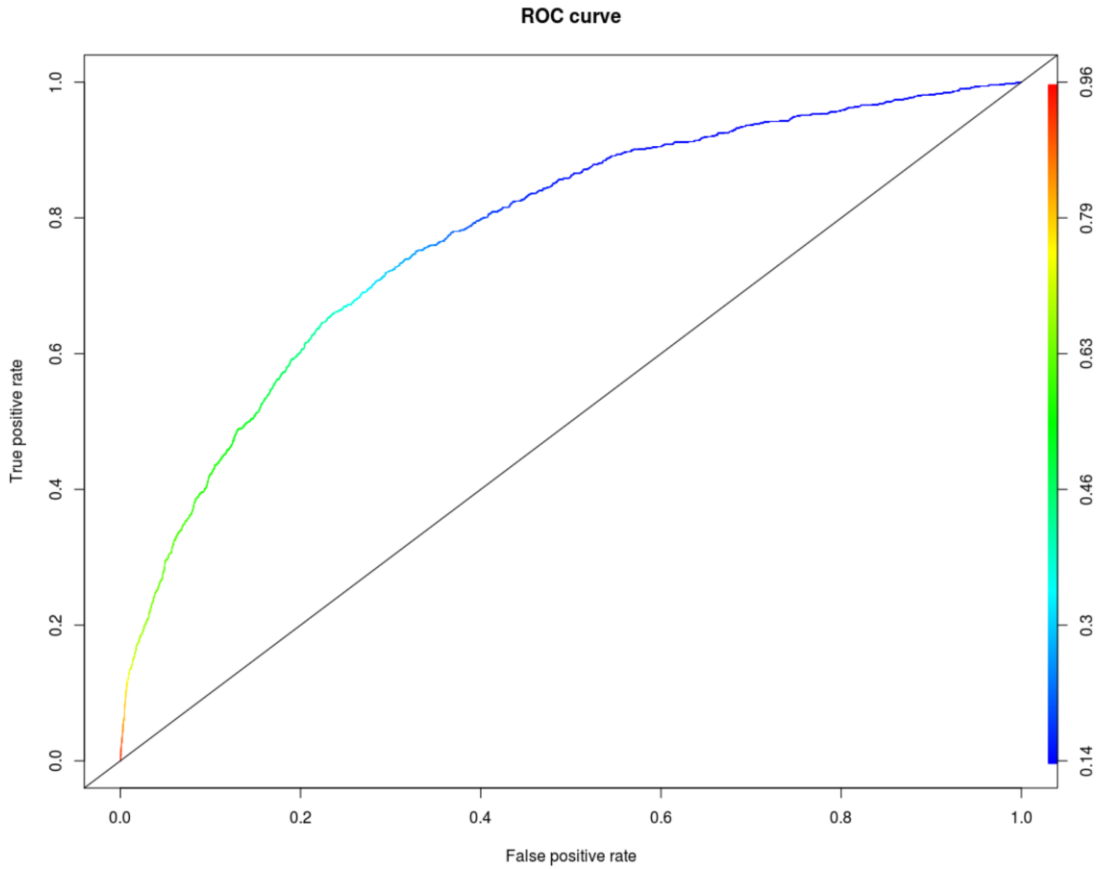
Embedding
(trainable dimensionality reduction)

CNN
bidirectional RNN

Clinically significant outcomes

eg
mortality
readmission
HbA1c





Initial results - 3 year runin analysis

Target: 1y mortality

AUROC 0.63

(current best performance 0.76)

input features	training targets	final output
demographic data	**mortality**	**optimum drug (per domain)**
numeric categorical TS	**hospital admission**	
numeric continuous TS	**HbA1c**	
categorical TS	**Blood Pressure**	
text	**BMI**	

what will the output look like?

Timing of information delivery

Ad hoc

- at time of consultations / other clinical contact
- on request
- via SCI diabetes frontend

‘Surveillance’ analysis: tackle therapeutic intertia

- batch analysis: custom group / clinic / population level
- ? 3 / 6 / 12 monthly
- lead to prompting of prescription change if seems beneficial
- ?ultimately opt-out / automatic prescribing approaches

Most granular

Detail of level of risk per domain (HbA1c / mortality etc)

Treatment options in detail – likely effect of each treatment on each domain

Best option drug / combination

Least granular / summary data / action points

Various levels of possible output at individual level:

	A	B	C	D	A+B	A+C	A+D
HbA1c	75	78	82	87	76	78	86
BP	140	135	154	143	134	132	145
BMI	26.4	28.2	27.3	31.3	32.2	28.2	24.2

HbA1c: drug A

BP: drug A+C

BMI: drug C+D

Mortality: drug A+C

use drug A + C

granular prediction detail

↓

domain recommendations

↓

single recommendation

CASE 2

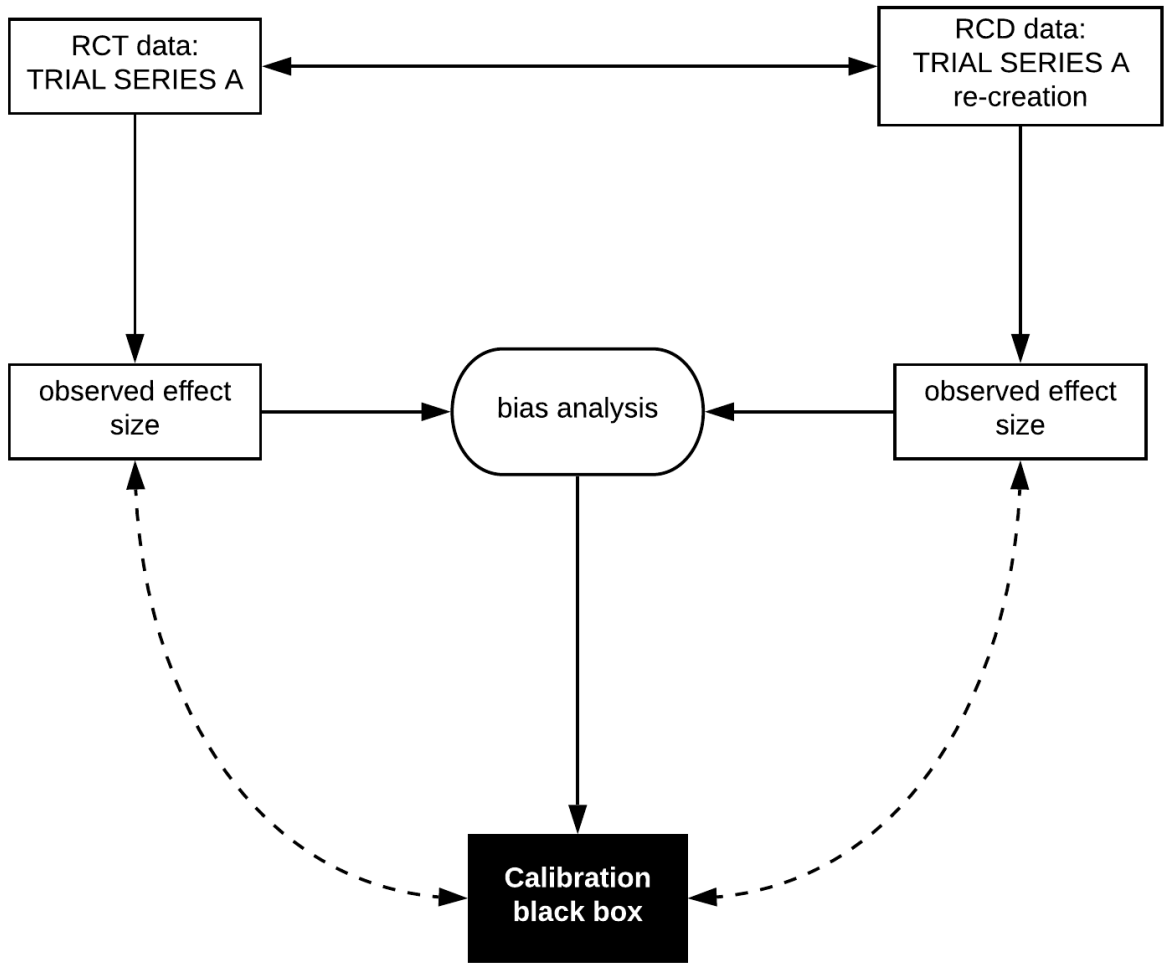
73 y male
Type 2 Diabetes - 3y duration
COPD
HbA1c 67

Case 2

clinical context slightly more complex
no good evidence to guide management
forced to extrapolate from multiple (siloed) disease area guidelines

but – many examples in the real world.
can we use the examples within RCD to explore treatment options, and relate this to current evidence

theme 2 **RCD / RCT calibration**
Extend in RCD to multimorbidity and map to RCT space



recreation of trials in real world data
exploration of bias and generation of calibration tool

A Comparison of Approaches to Advertising Measurement: Evidence from Big Field Experiments at Facebook*

Brett R. Gordon
Kellogg School of Management
Northwestern University

Florian Zettelmeyer
Kellogg School of Management
Northwestern University and NBER

Neha Bhargava
Facebook

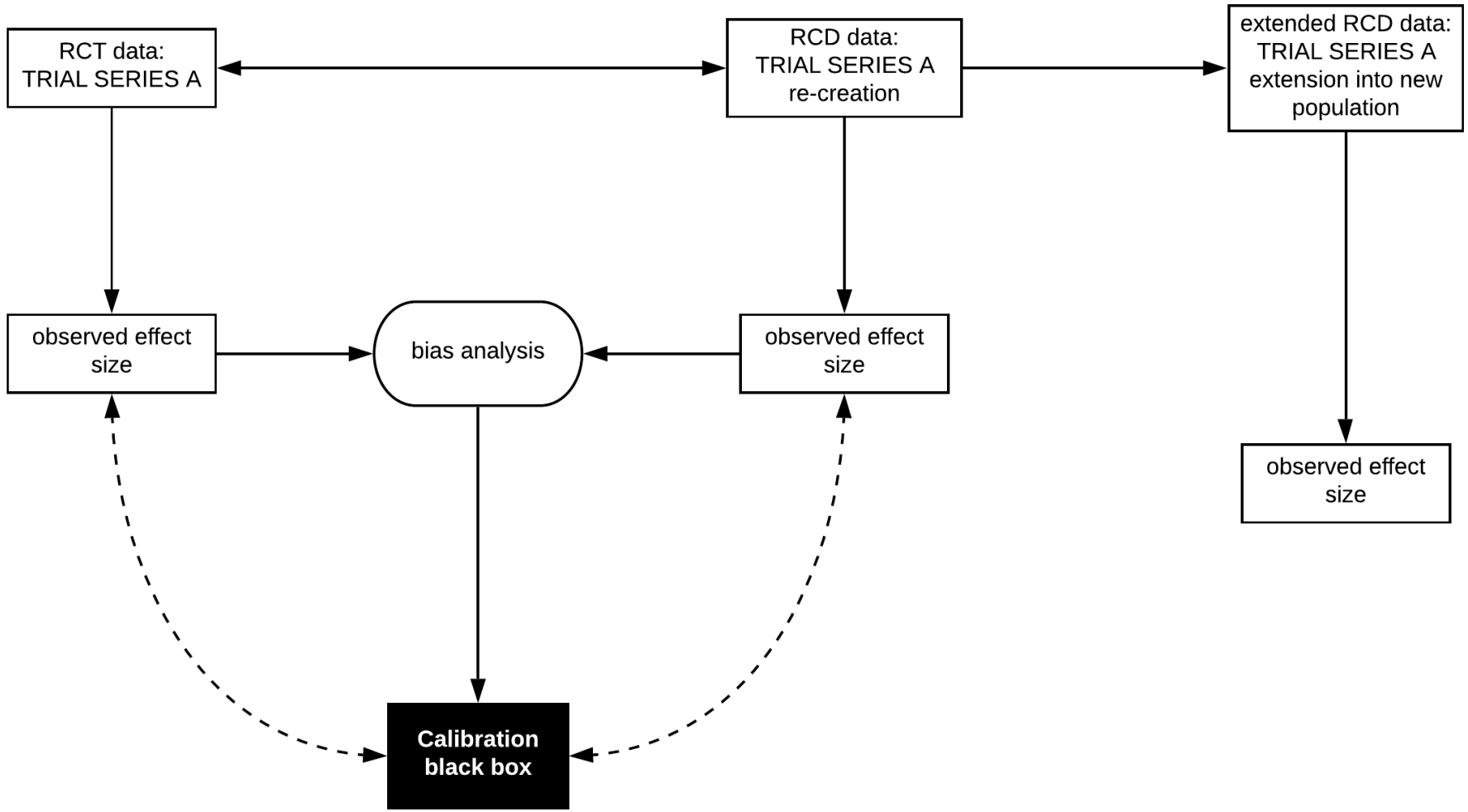
Dan Chapsky
Facebook

September 23, 2018

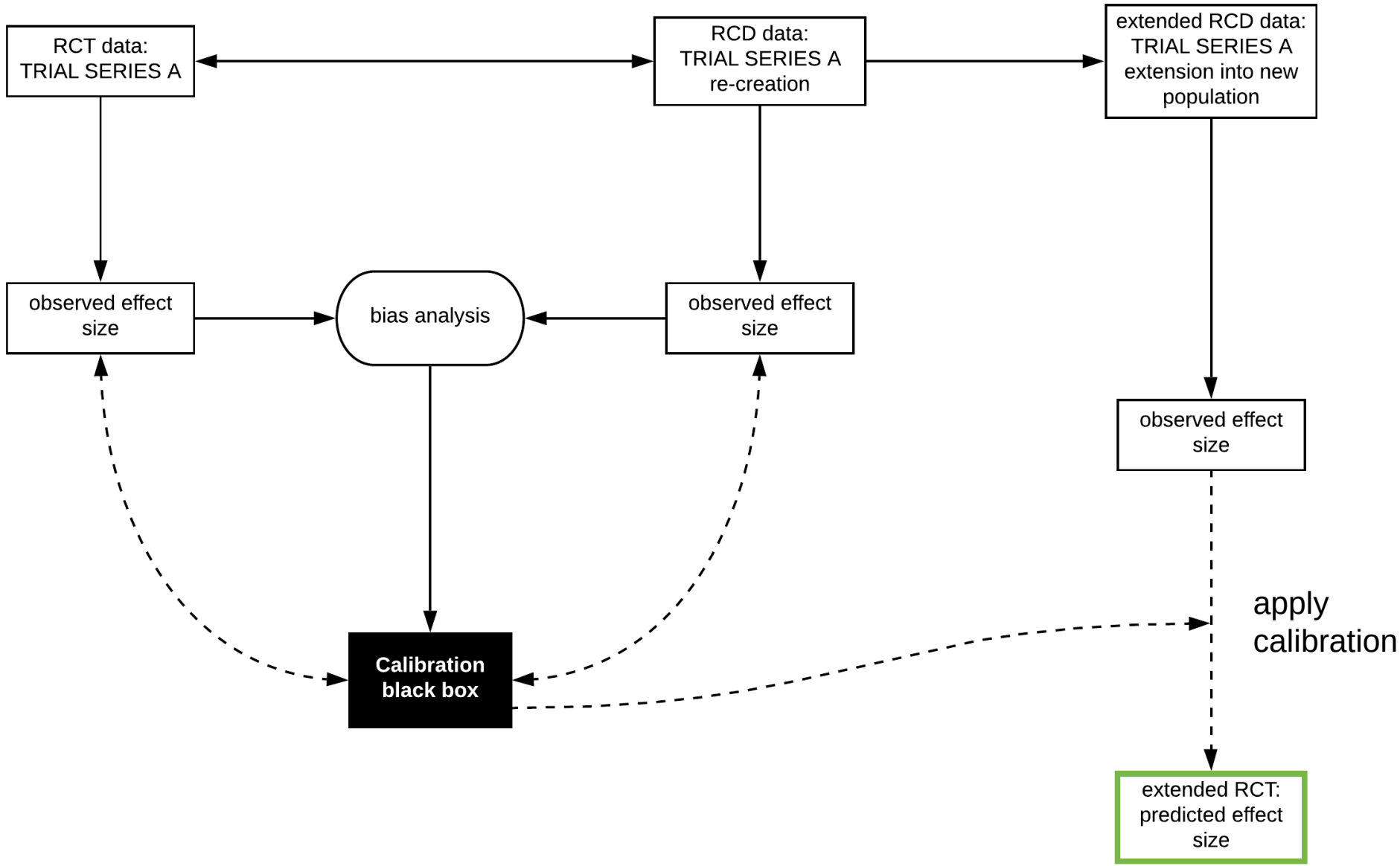
Abstract

Measuring the causal effects of digital advertising remains challenging despite the availability of granular data. Unobservable factors make exposure endogenous, and advertising's effect on outcomes tends to be small. In principle, these concerns could be addressed using randomized controlled trials (RCTs). In practice, few online ad campaigns rely on RCTs, and instead use observational methods to estimate ad effects. We assess empirically whether the variation in data typically available in the advertising industry enables observational methods to recover the causal effects of online advertising. Using data from 15 US advertising experiments at Facebook comprising 500 million user-experiment observations and 1.6 billion ad impressions, we contrast the experimental results to those obtained from multiple observational models. The observational methods often fail to produce the same effects as the randomized experiments, even after conditioning on extensive demographic and behavioral variables. In our setting, advances in causal inference methods do not allow us to isolate the exogenous variation needed to estimate the treatment effects. We also characterize the incremental explanatory power our data would require to enable observational methods to successfully measure advertising effects. Our findings suggest that commonly used observational approaches based on the data usually available in the industry often fail to accurately measure the true effect of advertising.

Gordon, Brett R., Florian Zettelmeyer, Neha Bhargava, and Dan Chapsky. n.d. "A Comparison of Approaches to Advertising Measurement: Evidence from Big Field Experiments at Facebook." *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3033144.



extension of trial into new (eg comorbid) population



apply calibration to express results in RCT space

Case 3

CASE 3

80 y female
Type 2 Diabetes - 20y duration
COPD, Psoriatic Arthritis, Ca Breast
HbA1c 67

clinical context much more complex
no good evidence to guide management
forced to extrapolate from multiple (siloed) disease area guidelines

this time – not so many examples in the real world.
? can we use generative techniques to develop multiple unobserved examples of the rare class

theme 3 generative techniques to augment RCD

related projects – synthetic datasets, rare class identification

requirements for synthetic data generator

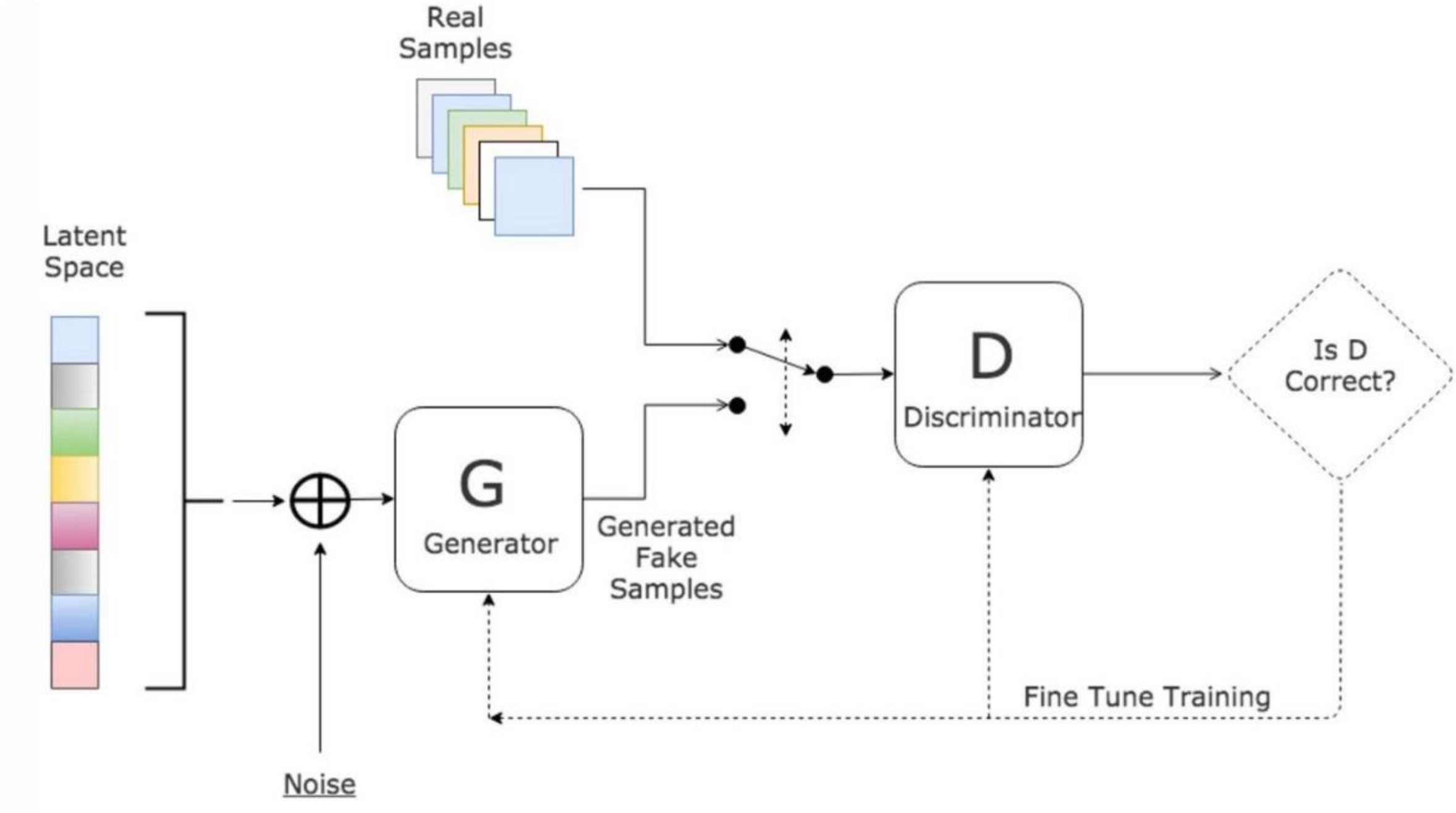
- generate multi-dimensional time series data
- reflect distributions of individual parameters
- include interactions/associations between parameters over time

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

Potential applications of synthetic data generator

synthetic data generation
- upsampling rare classes
- potential to help with another problem – data sharing

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


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

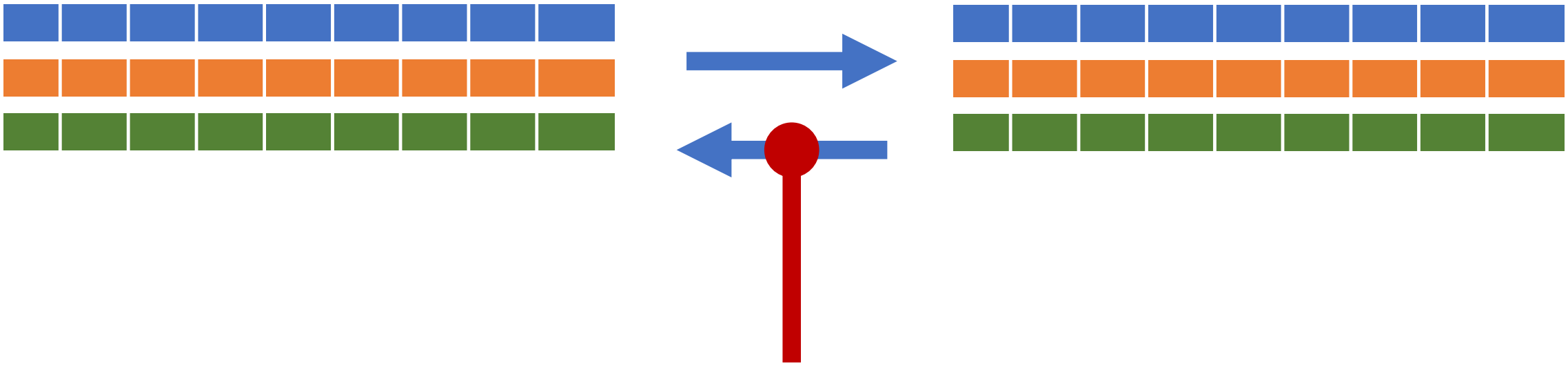
One Sentence Summary: Deep neural networks can generate shareable biomedical data to allow reanalysis while preserving the privacy of study participants.

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

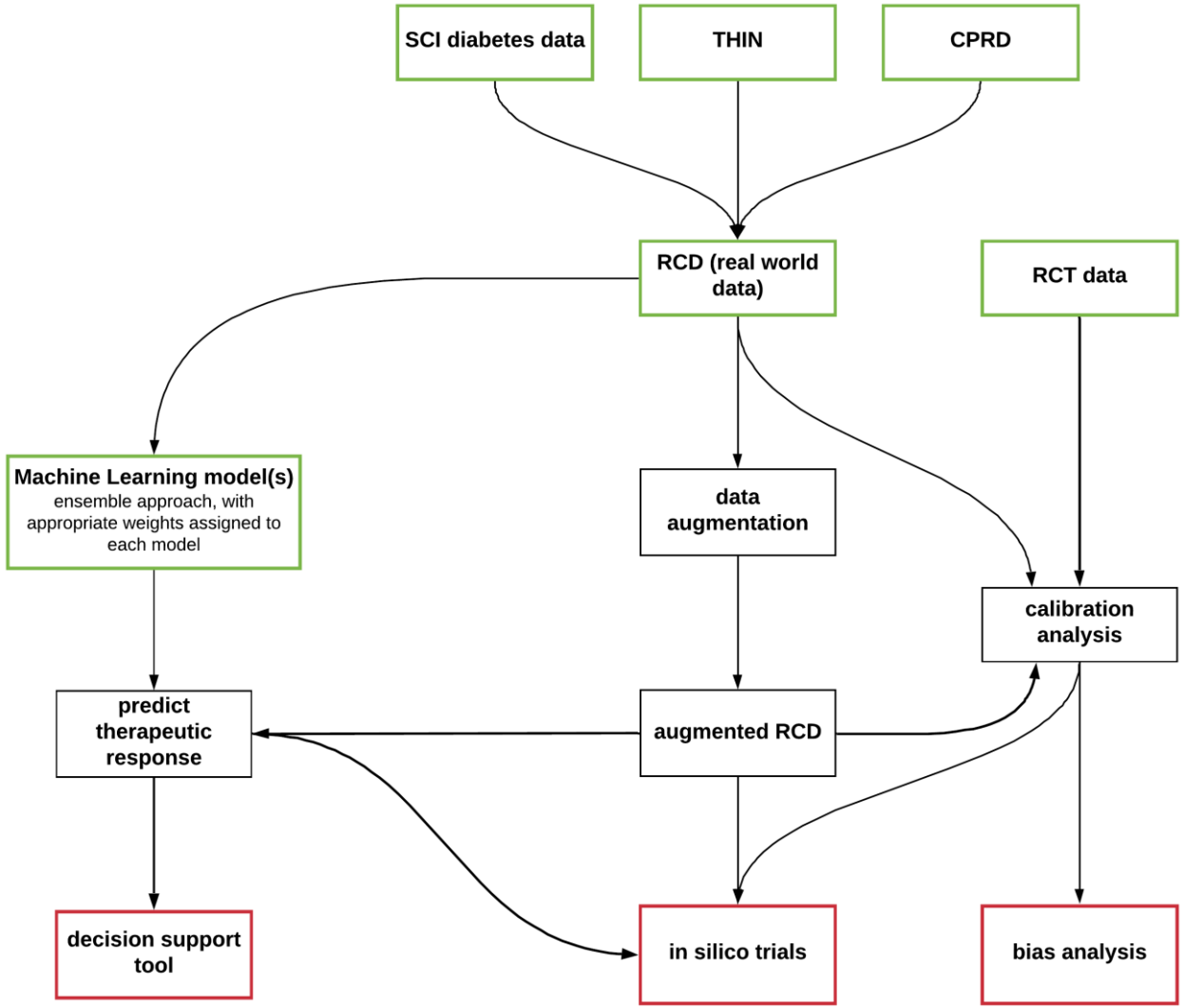
real time-series data

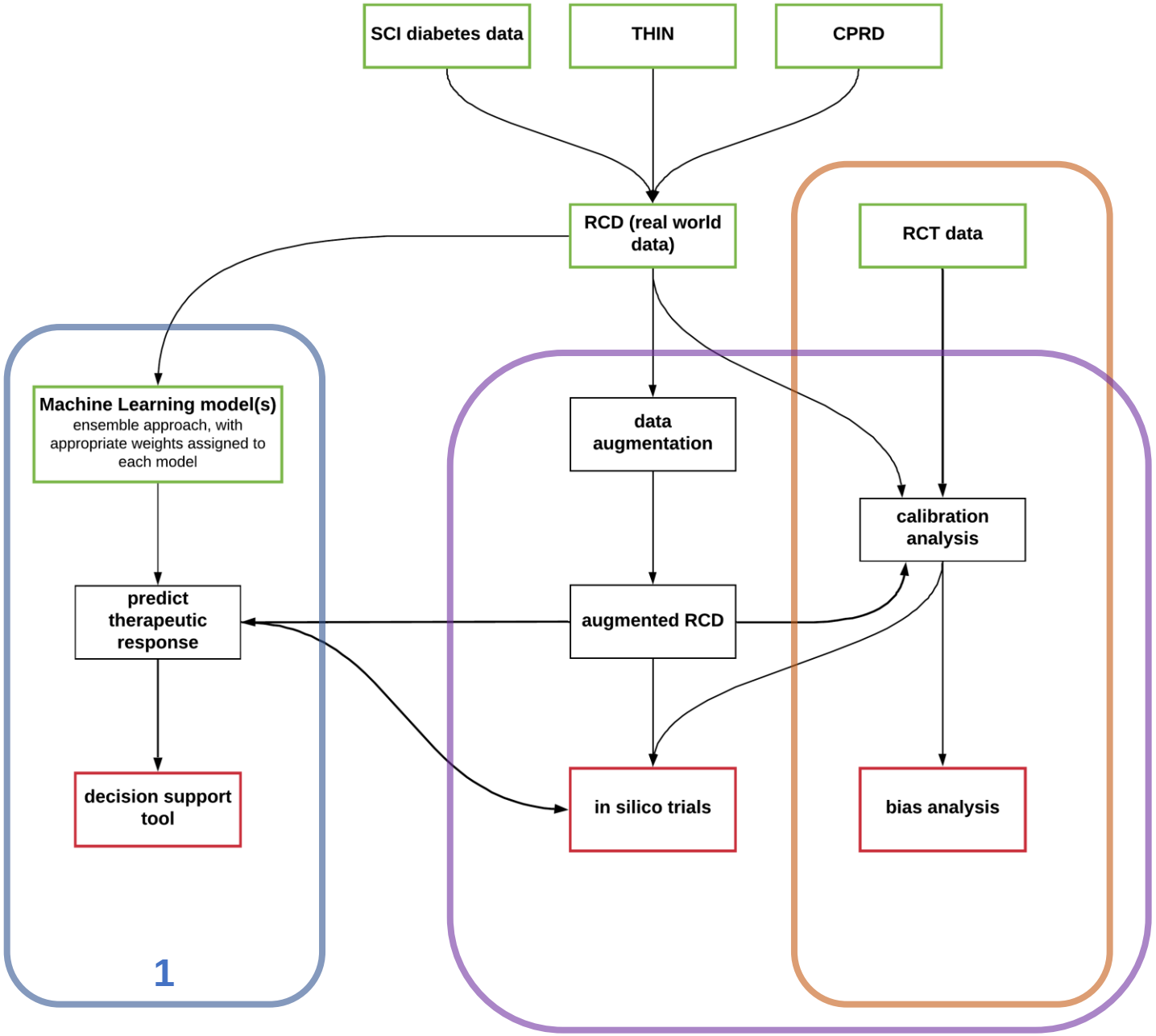
synthetic time-series data



differential privacy inhibits data leakage

- quality vs security tradeoff
- privacy budgets
- computational limits


SCI diabetes data
THIN
CPRD
RCD (real world data)
RCT data
Machine Learning model(s)
ensemble approach, with appropriate weights assigned to each model
data augmentation
calibration analysis
predict therapeutic response
augmented RCD
decision support tool
in silico trials
bias analysis



Theme 1

Drug response prediction

Largely application / implementation problem

Theme 2

RCD / RCT calibration / bias analysis

Application + methodological problems

Theme 3

Generative techniques / synthetic data

Largely methodological

predict the response (effect size within a specified domain) to an **arbitrary intervention** in an **arbitrary population** (or individual)

report effect size in both RCD and RCT-space

- predict optimum therapy choice for an individual / population (decision support)
- predict at-risk populations for adverse events (decision support)
- investigate potential for indication expansion for existing therapies (trial design)
- Investigate sources and size of hidden biases

glucose.ai group, NHS GGC
Dr Greg Jones
Dr David Carty
Dr Jim Boyle
Dr Debbie Morrison
Dr Sean Harbison

myDiabetesIQ
Dr Debbie Wake
Dr Scott Cunningham
Dr Felix Agakov
Dr Nicky Conway
Dr Sam Philip

HDR UK Midlands
Prof Alastair Denniston

kvatchii
Prof Sandosh Padmanabhan
Ajay Sharma
Dr Shyam Visweswaran

redstar
Andrew Conkie

thinkingAI group, University of Birmingham
Dr Krish Nirantharakumar
Krishna Gokhale
Dr Malcolm Price
Dr Chris Yau
Dr Tom Taverner
Prof Peter Tino