Prediction and Risk Scores

Tom Ron • ML4HC • April 2019

Agenda

- Biases in electronic health record data due to processes within the healthcare system: retrospective observational study
- Meaningless comparisons lead to false optimism in medical machine learning

Biases in electronic health record data due to processes within the healthcare system: retrospective observational study

Denis Agniel, Isaac S Kohane, Griffin M Weber

“The hour of the day the test was ordered, the day of the week, and the amount of time between consecutive tests is more predictive of three year survival than the actual value of the test result, for most tests”

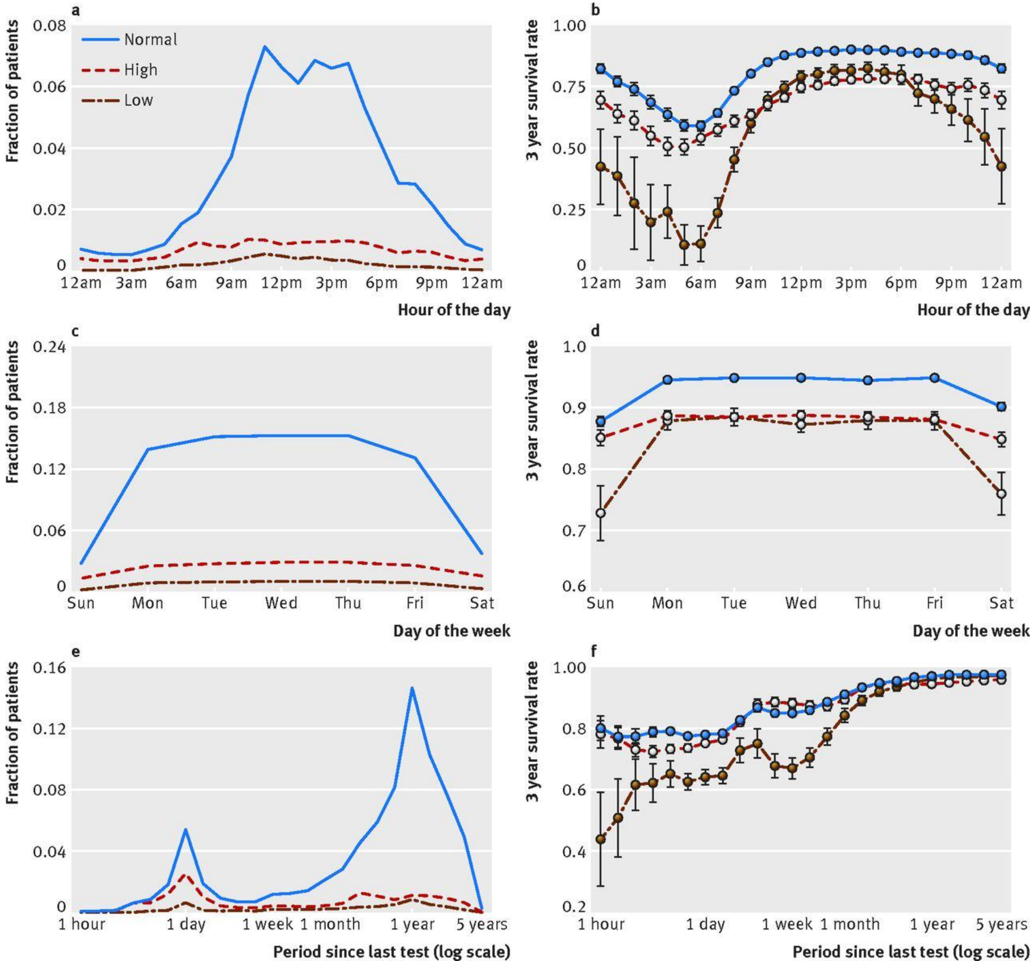
Experiment

- **Goal**: predict 3 years survival
- **Data**: 8.8M observations (670k patients) of 272 laboratory tests from 2 hospitals were used in the experiments.
- **Features**: patients' pathophysiology (value + high \ low flag) and healthcare process dimensions (hour of day, day of week, previous test) of a single laboratory test observation

Experiment

- First experiment - logistic regression + Age, Sex, Race (ASR) + test presence
- Second experiment - using GAM with logistic link, features ASR + …

Results



a
Fraction of patients
0.08
0.06
0.04
0.02
0
Normal
High
Low
12am 3am 6am 9am 12pm 3pm 6pm 9pm 12am
Hour of the day
b
3 year survival rate
1.00
0.75
0.50
0.25
0
12am 3am 6am 9am 12pm 3pm 6pm 9pm 12am
Hour of the day
c
Fraction of patients
0.24
0.18
0.12
0.06
0
Sun Mon Tue Wed Thu Fri Sat
Day of the week
d
3 year survival rate
1.0
0.9
0.8
0.7
0.6
Sun Mon Tue Wed Thu Fri Sat
Day of the week
e
Fraction of patients
0.16
0.12
0.08
0.04
0
1 hour 1 day 1 week 1 month 1 year 5 years
Period since last test (log scale)
f
3 year survival rate
1.00
0.8
0.6
0.4
0.2
1 hour 1 day 1 week 1 month 1 year 5 years
Period since last test (log scale)

Results

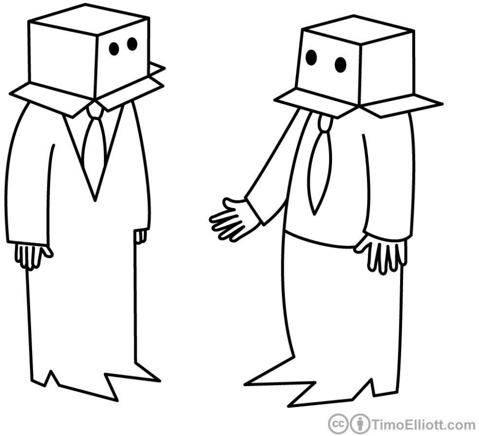
- The presence of a laboratory test in a patient's record, regardless of any other information about the test result, has a significant association with the odds ratio of death in 233 of 272 (86%)
- For 60% tests including both patient pathophysiology and healthcare process variables in the models is better than using patient pathophysiology or healthcare process alone.
- The time interval between consecutive tests is the single most predictive variable for 76 of 210 (36%) tests, followed by the value of the test result in 56 (27%) tests, and the hour of the day in 47 (22%) tests
- 30 day readmission as the outcome measure, rather than three year survival, and found similar results.

Critic

- PoC - not clear how to go on from here
    - Severity score
    - Models with multiple tests
    - Anomalies and Fairness
    - Model changes in guidelines \ dynamics
- 3 years survival - not the doctor objective function

Takeaways

- Think of features out of the box -
  - Who prescribed the test
  - Room in unit
- Synergy between clinician and ML researchers \ practitioners



"What a coincidence!
I'm finding it hard to think out of the box, too!"

Meaningless comparisons lead to false optimism in medical machine learning

DeMasi O, Kording K and Recht

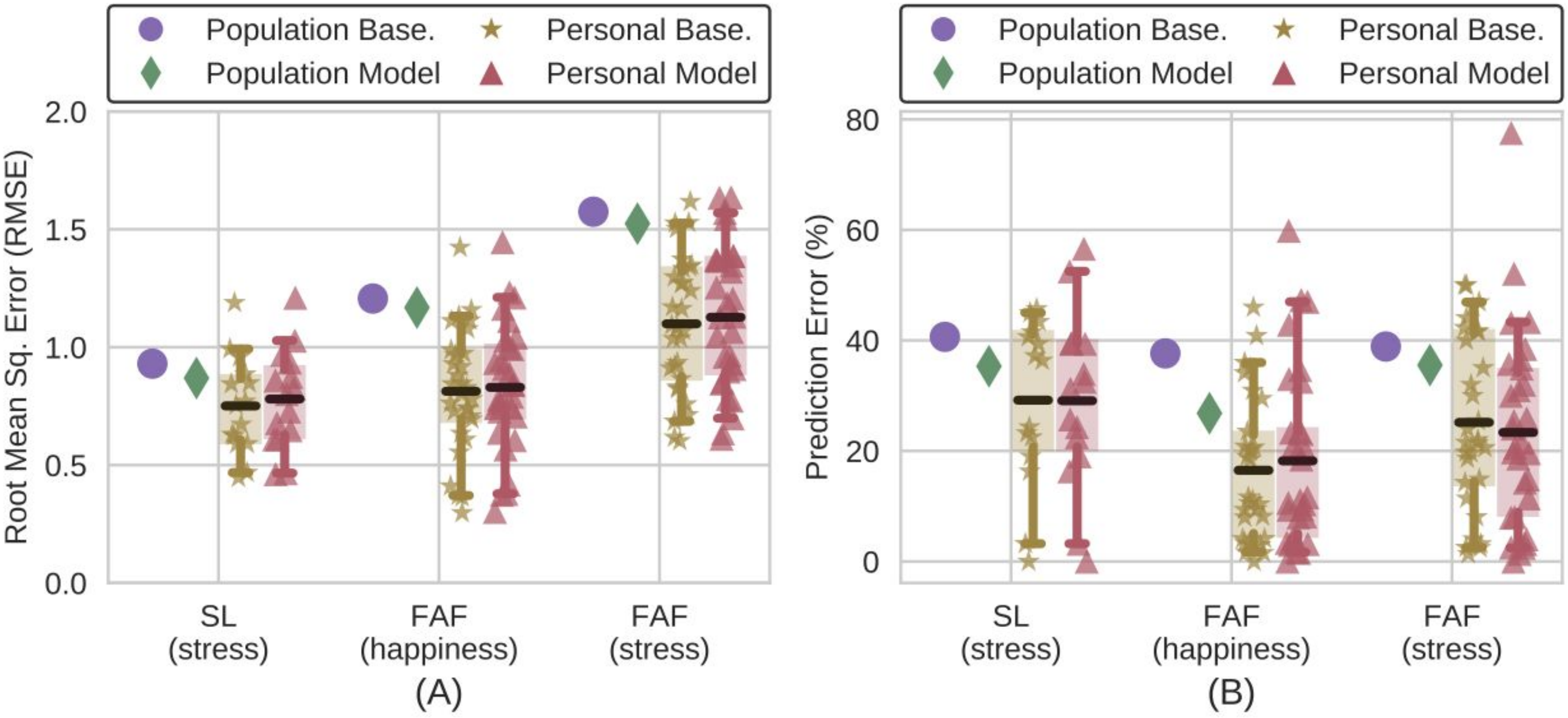
“mental health conditions need long-term monitoring and clinical monitoring is expensive, but automatically tracking a user with ubiquitous sensors is cheap.”


מה אתם הייתם עושים?

Terminology

Personal baseline - each individual is at a constant state, but that state can differ between individuals.

Population baseline - all individuals are always at the same state.



Population Base.
Personal Base.
Population Model
Personal Model
Root Mean Sq. Error (RMSE)
2.0
1.5
1.0
0.5
0.0
SL (stress)
FAF (happiness)
FAF (stress)
(A)
Prediction Error (%)
80
60
40
20
0
(B)

Experiment

Goal -

- Predict happy or stressed or not on a given day
- Predict average level of happiness or stress that a participant reported on a given day

Data - StudentLife and Friends and Family

Experiment

Preprocessing

- Full clustering - fit GMM to all locations for each participant to identify locations frequented by participants. Max 20 clusters. Home and work definitions.
- Stationary clustering - K-means clustering on stationary points only.

Features

- Fraction of a day participant is not stationary
- Log likelihood of a day from the GMM to estimate how routine the day was
- # GMM clusters visited in a day

User lift

User lift - the difference of the personal model with the personal baseline

Dataset	Problem	Model	Avg. Personal Baseline Error	Avg. Personal Model Error	Avg. User Lift (Error)	p-value
SL - Stress	binary	Log.Reg.	29.19%	29.09%	**0.10**	.481
FaF - Happiness	binary	SVM(rbf)	16.51%	18.67%	**-2.17**	.967
FaF - Stress	binary	SVM(rbf)	25.17%	23.35%	**1.82**	.240
SL - Stress	regression	Elastic Net	0.75	0.78	**-0.03**	.988
FaF - Happiness	regression	Elastic Net	0.81	0.83	**-0.02**	.999
FaF - Stress	regression	Elastic Net	1.10	1.13	**-0.03**	1.000

Limitations

- Study cohort is not clinical population
- Sample size is small
- Study duration is limited

Literature review

Find relevant literature

Choose papers for literature review

Establish a baseline

Extracting baselines form the different papers

Identify error of baseline and best ML

Extract results from the different papers

Find relevant literature

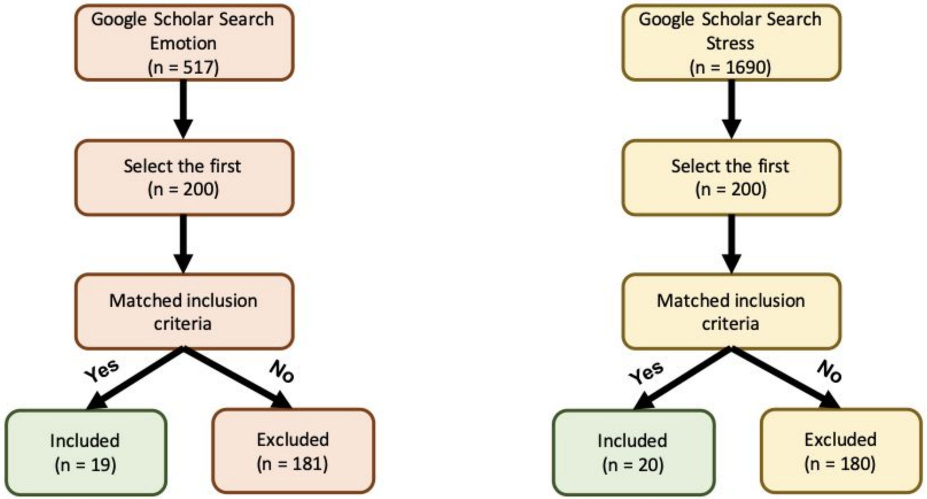
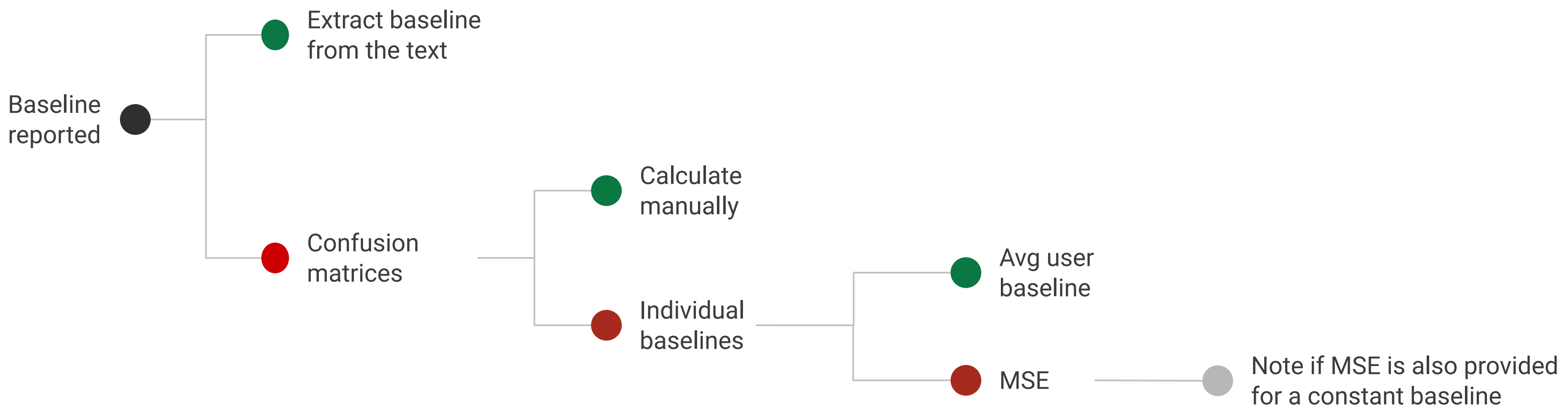


Fig 2. Diagram of literature review process.

Establish a baseline



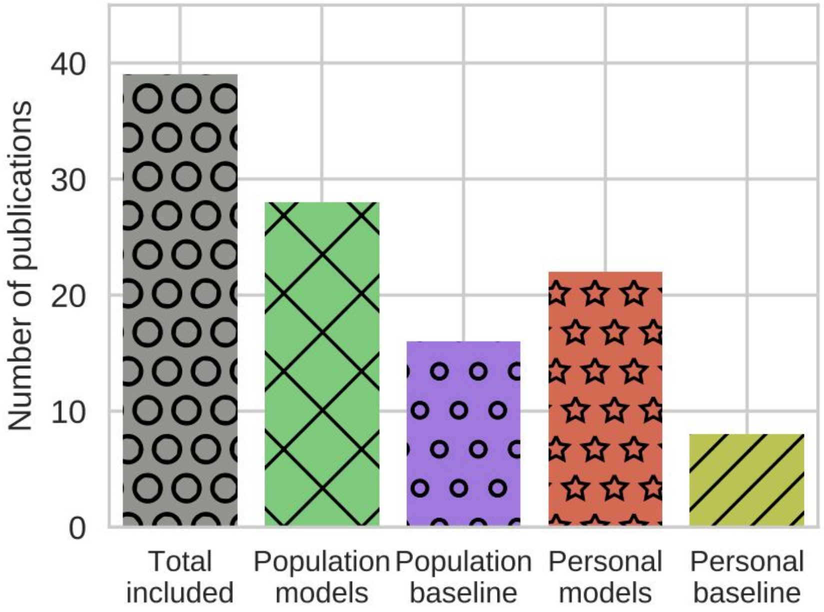
Identify error of baseline and best ML

Model prediction error for multi-class classification

- Results are broken for personal models by individuals → average
- Accuracy for multiple objectives → best result for each objective
- Multiple feature sets and models → best performing model
- Folds in cross validation scheme - ignored.
- Uniform baseline based on classes probability

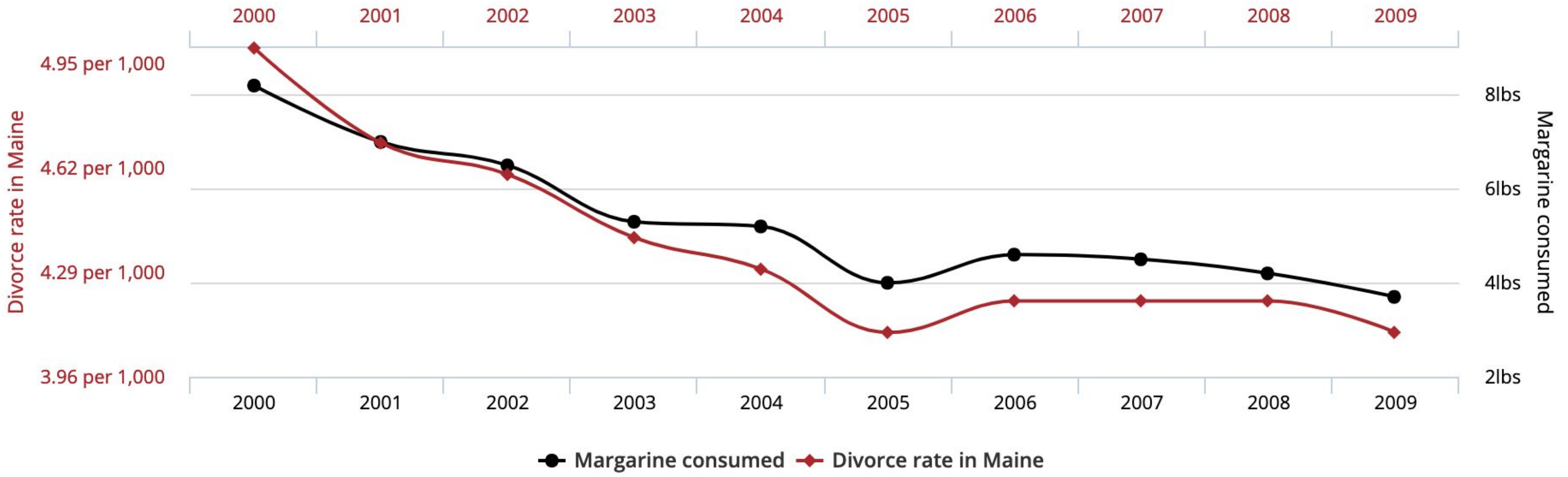
Literature review

- 77% of the publications reviewed compared to population baselines.
- When personal baselines are reported algorithms often add little or nothing



Divorce rate in Maine
correlates with
Per capita consumption of margarine

Correlation: 99.26% (r=0.992558)



tylervigen.com

Data sources: National Vital Statistics Reports and U.S. Department of Agriculture

Takeaways

- Read with critical eyes - sample size, duration, clinical population, can the result be generalized, reproduced.
- Results should be meaningful
- Gap between theory and ready to deploy model


That's all Folks!

References

Agniel D, Kohane IS, Weber GM. Biases in electronic health record data due to processes within the healthcare system: retrospective observational study. BMJ 2018;361 doi: 10.1136/bmj.k1479

DeMasi O, Kording K and Recht B (2017) Meaningless comparisons lead to false optimism in medical machine learning. Ed. Y-K Jan. Public Library of Science PLoS ONE 12, e0184604–15

Wang R, Chen F, Chen Z, Li T, Harari G, Tignor S, et al. Studentlife: assessing mental health, academic performance and behavioral trends of college students using smartphones. In: Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing. ACM; 2014. p. 3–14. 26.

Aharony N, Pan W, Ip C, Khayal I, Pentland A. Social fMRI: Investigating and shaping social mechanisms in the real world. Pervasive and Mobile Computing. 2011;7(6):643–659.