AI for decision support in Health – how to make it work

Mark van Gils, Ph.D., Adjunct Professor, Principal Scientist, Smart Health
mark.vangils@vtt.fi
Goal of this tutorial

- Get an understanding of why AI applied to healthcare is not 'business as usual'

- Get acquainted with the main practical problems and understand how to address them
"generic" AI, machine learning, signal processing

Healthcare specific data

messy

"challenging"

complex

Healthcare environment and processes

Healthcare specific targets
Personalised care: health care treatments to match the unique characteristics of individuals. Genomic and other ‘omics data with other health and behavioral data including data from sensors. **AI is needed to process huge amounts of data.**

Automated health data analytics: Automated analysis of complex health data - imaging, electronic health records, sensor data – reliable quantification and interpretation. **AI to pre-process, and analyse data.**

Continuous citizen-centric care: Improve continuous preventive management of health of individuals by automatically monitoring and integrating information. **Use AI to analyse, interpret changes in health status. Engage. Motivate.**
The Data – measured from humans

- Every person is different (their own reference) [inter-subject variability]
  - What is a low heart rate for me, could be normal for you

- Your personal ‘normal’ may change over time [intra-subject variability]
  - My blood pressure today may be higher than yesterday, it doesn’t mean I suddenly need to see a doctor (nervousness for giving a presentation? coffee usage? Short night sleep?)

- How to deal with these variations in decision making?
Strategy A – Just use all data as one batch

- Ignore any individual variation, just add all the data in one big batch
- Motto: "The AI will sort it out itself if we have enough data"

- Might work – IF we have lots of representative data. Not optimal though (and difficult to sell to clients (clinicians))
Strategy B – personalisation, stratification

- Use statistical methods to account for variation in different subjects, e.g. use partial correlations, patient ID’s as dummy variable

- Normalise all data to personalised baseline values, MinMax(); StandardScaler - Gaussian with 0 mean and unit variance

- Stratification, binning into groups of similar persons (based on age, gender, education,…)

- Correct for ’nuisance variables’ such as age (e.g. age has an effect on brain size, regardless of diseases like dementia)

- Dedicated methods like histogram transform
Adaptive centile correction algorithm. Each feature is remapped according to the feature-wise centile curve (yellow) to the age of the subject (green vertical line). The feature-wise centile is computed as a weighted average from the nearest centile curves acquired with LMS method with Yeo-Johnson power transformation.
Histogram transformation example: variability of heart rate between patients

HR intra-patient
Two narrow distributions

Pat 34  Pat 45

Adult Mean

HR intra-patient
A broad distribution

Pat 55

Adult Mean
How do we create an individual normalisation transformation?

estimate group distribution

estimate individual distribution
data either
  1. since the beginning of recording, or
  2. over running time window

calculate combined distribution
A \times \text{group} + B \times \text{individual}
(eg A=0.7, B=0.3)
calculate cumulative combined distribution

>> histogram transformation
Missing, incomplete data

- In multi-variate/multi-modal data sets
- In time series

- Most classifiers don’t work correctly when data input vectors have empty elements or NaNs – what to do?
Dealing with Missing data

- In the majority of applications where we combine lots of modalities (images, -omics, text, patient monitors), some elements are missing from the input vector

- Only use complete cases?
  - Leaves only a fraction of the data, we throw away too much valuable data

- Impute all missing data? With averages, interpolations etc.
  - May work sometimes, but defeats eg the purpose of clustering

- Compromise: Only use cases that have at least eg 70% of all elements present, impute the rest

- Consider data analysis methods that may work with incomplete data, eg c-means clustering, DSI*

*DSI: Disease State Index, Mattila et al
Missing data in time series

- Imputation (splines, regressions etc), within reason (and in agreement with domain experts).
  A curve that is visually pleasing is not necessarily clinically valid.

- **BUT NOTE**: the fact that there is missing data is information in itself

- For example, decreased frequencies of self-measurement may indicate changes in the person’s state
Weight loss took place during periods of daily self-weighing, whereas breaks longer than one month posed a risk of weight gain. Missing data in weight management studies with a weight-monitoring component may be associated with non-adherence to the weight loss programme and an early sign of weight gain.
From model-based to black-box approaches

- Rule-based approaches
- Mathematical models
- Physiological models
- System dynamic models
- Probabilistic models
- Syntactic pattern recognition
- Decision trees
- ...

- Black-box approaches
Black-box approaches?

- Traditionally, the argument has been that if the functioning of an algorithm is not explainable it cannot be used for making decision support in healthcare.

- Explainability does improve confidence in decision making, and is HIGHLY VALUED by healthcare professionals.

- But, its strict necessity may be decreasing for certain tasks, especially the more routine subtasks.
  - E.g. a CNN that segments images very well does not necessarily need explainability.
Results presentation and interaction

- Patient data can be very high-dimensional
- We as data scientists love to find clusters and visualise interesting structures in the data
- Tools: PCA plots, t-SNE, SOMs, Sammon mappings, ….
Clinicians’ preferences

- Overall, high-dimensional visualisations and (non-)linear mappings are difficult concepts to communicate to non-data scientists.

- In addition, clinicians are often conservative, they prefer:
  - Raw data, time series – old fashioned timelines that they are used to seeing
  - Quick overview + details on demand
Clinical Platform for decision support in dementia

EU projects: PredictND, VPH-DARE@IT
Assessment of performance

- We need to measure how well our AI method does

- Commonly used measures:
  - Accuracy, sensitivity, specificity, positive prediction probability (PPV), negative prediction probability (NPV)
  - Precision, Recall, F-score (=harmonic mean of precision and recall)
  - ROC curve and AUC-ROC

*Note: Precision and recall are related to PPV/NPV and sensitivity respectively if we have a 2-class problem*
Alarm overload

1. Alarm Hazards & 2013 & 2012
2. Infusion Pump Medication Errors
3. CT Radiation Exposure in Pediatric Patients
4. Data Integrity Failures in EHRs and other Health IT Systems
5. Occupational Radiation Hazards in Hybrid ORs
6. Inadequate Reprocessing of Endoscopes and Surgical Instruments
7. Neglecting Change Management for Networked Devices and Systems
8. Risks to Pediatric Patients from “Adult” Technologies
9. Robotic Surgery Complications due to Insufficient Training
10. Retained Devices and Unretrieved Fragments

<table>
<thead>
<tr>
<th>Clinical</th>
<th>Overall rate</th>
<th>Relevant Alarms</th>
<th>False Alarms</th>
<th>Nuisance Alarms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Priority</td>
<td>(/hr)</td>
<td>%</td>
<td>%</td>
<td>%</td>
</tr>
<tr>
<td>ECG</td>
<td>1</td>
<td>0.7</td>
<td>9</td>
<td>41</td>
</tr>
<tr>
<td>SpO2</td>
<td>2</td>
<td>0.9</td>
<td>5</td>
<td>63</td>
</tr>
<tr>
<td>InvPress</td>
<td>3</td>
<td>0.6</td>
<td>20</td>
<td>27</td>
</tr>
<tr>
<td>ImpResp</td>
<td>4</td>
<td>0.4</td>
<td>17</td>
<td>45</td>
</tr>
<tr>
<td>Other mon</td>
<td>5</td>
<td>0.2</td>
<td>19</td>
<td>22</td>
</tr>
</tbody>
</table>

Prevalences (percentage of disease cases in population)

- Health data are often highly imbalanced. With the number of disease cases (luckily) much smaller than the number of healthy cases.
- PPV (or, Precision) is especially intuitive for health tests: "If the test says that I have a disease, what is the probability I actually have that disease?"
- `sklearn.metrics classification_report()` gives precision (PPV) as one of its main outputs.

However, precision (or, PPV) is dependent on prevalence (next slide).
Hence, we often prefer sensitivity, specificity, confusion matrices, ROC curves.
Using the same test in a population with higher prevalence increases positive predictive value. Conversely, increased prevalence results in decreased negative predictive value.

https://onlinecourses.science.psu.edu/stat507/node/71
In other words

- If I want to make a classifier to detect a "rare" disease, e.g. with prevalence 1% (1 in 100 cases is 'disease', 99 are 'healthy')

- And I use some scheme to have a training & test set that has 50% of the cases as disease and 50% as healthy, to make my "training easier"

- The PPV values that are reported after cross-validation may be enormously inflated vs what I will encounter in real-life use

- Hence: rely more on confusion matrices, sens, spec
Cross Validation

it is common practice to divide the total data set into 2 parts: the cross-validation (CV) set (this is what we work with) and the test set. The test set is left absolutely untouched during the entire development process, only when we have finalised our classifier we can use it for performance assessment.

the cross validation set is again split into 2 parts; training and training-test set. the patterns that fall into the training and training-test sets are ‘circulated’

keep far away from the development process!!

Requires long-term view and self-discipline!!
Complex cases, co-morbidities

- Most people (at older age) have multiple diseases at the same time (diabetes & cardiac disease; dementia & high blood pressure, etc)
- This makes classification more difficult, as we have no clear class 0 vs class 1 data anymore, but more complex mixes
- This is enormously tricky – for clinicians as well as data scientists, but one of the Grand Challenges to solve
Cost functions for training an AI

- $\text{loss} = \text{tf.losses.mean_squared_error}(...$
- $\text{loss} = \text{tf.losses.softmax_cross_entropy}(...$
- Etc..

Traditionally we have been using RMSE or other convenient mathematical functions on straightforward data.

- Other measures might be more relevant
- Clinical relevance? Confidence? Speed? Quality of Life? Financial cost?
- Moving towards reinforcement learning?
Incorporating AI into real clinical practice takes considerable efforts, but we will succeed

- If we keep in mind

  - Things like data collection and curation take a long time

  - Your toolbox is big, think not only deep learning, but also old-school statistics, linear filters, logistic regression, … – they can serve you very well. Try to see the bigger picture and use all tools that may be useful

- Co-operation: data-driven and modelling approaches

- Healthcare professionals are your friends, users and guides: visit and discuss with them – often