11 prediction machines and social justice: the end of population health?

The pure technical scoring of predictive power as a measure of the accuracy of a prediction (using ROC/AUC, Receiver Operating Characteristic / Area Under the Curve) may conceal design problems most notably in the AI training data: garbage in / hallucinations out....
This relates to the extent to which these models are learning on data that takes account of the wider social context in which ill-health sits. It is to be determined whether AI models are trained to embody any particular social reality and not just the extent to which they deal with identification of individuals within defined clinical parameters which may include forms of bias (caused by the trainers and by the selection of the data itself) and discrimination (from the way it works).
The concern here is that the use of machine learning is not just an issue of assessing models for their technical accuracy, but of understanding the extent to which the use of these models is compatible with our notions of 'social justice' and the fact that humans are individuals and one cannot simply generalise as machine learning training does.
As they come into wider use, we should expect compliance on basic principles of equity and equal treatment of individuals to surface as critical design features, not just the ability to identify and work with cohorts as in population health.
But since these are really black boxes with few ways to peak inside, the consequences of negligence in design must be addressed.

Selection of data by the model designers is a source of error whether deliberate, accidental or carelessness, and may lead to unacceptable social bias or unfair allocation of clinical resources if this training causes 'learned bias'. We have seen this in the recent past on the flaws in face recognition software. We also have the socially learned bias in humans.

It should be unacceptable that prediction models might be inconsistent with social values. To be specific, we would not want machine learning models to produce unfair allocation of healthcare resources; for example to prioritise over-serviced individuals because the machine learned on information that biased it toward that type of individual. These people may also be over-represented in the training data in virtue of their overconsumption of healthcare.
But do we allow humans to have a wider range of unacceptable bias than prediction machines? The human excuse is that we should know better, but don't while the machine is just built that way, but biased humans.
Think about it this way. In some countries, there are those who are under-serviced, and left-behind when it comes to access to healthcare and treatment, and their data may not be in training databases. From the perspective of the machine, it won't ask for the missing data and won't know it is missing.
That means that these people are 'invisible' to the model: this doesn't mean the model is 'blind', it means the model is biased. Administrative and clinical databases are full of errors, and missing people that arises from systemic discrimination in the distribution of healthcare resources and facilities so that some communities are healthcare deserts while others have massive excess and duplication of facilities.
Our personal belief in our own infallible judgement suggests that this is the most important source of bias in clinical reasoning. When compared to machine learning, we see a higher level of accuracy in the predictions from neural networks than from people.
If the prediction machine is more accurate than unaugmented / unaided clinician judgement, on what basis can clinicians override these systems? It is to be assumed that a clinician over-

riding a prediction suggestion for patient treatment would be correct, but for the possibility that the clinical guidelines are biased, the doctors is biased, or the disease model is biased.
All this highlights the importance of including significant assessments of bias through process of developing machine intelligence. Where care institutions with public mandates (publicly funded) are concerned, they need to be vigilant of social justice considerations in the use of machine learning. This may be challenging if their business model involves chasing reimbursements and responding to cost-efficiency criteria, and not prioritising effectiveness and value for money outcomes.
Individuals may have had experience of prediction machines from applicant profiling by medical insurance, and arbitrary denial of coverage. Is it possible that this is what is happening in the US with Medicare Advantage? Hmmm.

Profiling (patterns), which is a core competency of prediction machines, can result in a variety of undesirable consequences when put into the wider social context. The problem of profiling is embedded in the design (training and algorithms). And this design of necessity captures how we value people under different (predicted) service/risk models (patterns), e.g. homeless, rich, etc. And we're back to invisible people.

The maldistribution of healthcare resources is a known problem in healthcare, with over and under serviced populations, exacerbated by articulate individuals who are convincing in their need to over-consume healthcare resources, and by silent individuals who are unable to claim access.
The ethical question being asked is 'are those people identified for specific treatment though prediction machines more worthy of special consideration than those not identified'? Or, is every person entitled to equal access to unequal resource distribution by a prediction machine that may identify some people as more at risk and therefore more deserving of greater resource allocation?
What would John Rawls say?
In effect, can prediction machines give voice to those who need healthcare resources in way that produces a fair allocation based on manifest need rather than a formulaic distribution of resources.

The issue looks like this:

Prediction machine matrix for the distribution of benefits and risks/costs [after Wilson]		
	High predicted benefit: treatment is efficacious	**Low predicted benefit: treatment is not (very) efficacious**
High predicted individual (health) risk	N-of-1 has the risk and will benefit from treatment (personalised, individualised allocation of resources (not cohorts)	there is risk but the treatment benefits are low
Low predicted individual (health) risk	there is benefit but the risk is low	Do not provide

There are serious justice considerations embedded in this overview table: the diagonal summarises the fundamental healthcare debate about how much resources to allocate and for what purpose. Where we want to be is in the N-of-1 box. The diagonal otherwise represents misallocation of healthcare resources (worried well, articular middle class, discrimination, mal-distribution of access, etc.).
As an analytical model this also shows why prediction models reframe healthcare priorities based on individualised precision care and not population-based epidemiology. Individuals are their own normal distribution with the standard deviations being the risk boundaries in their personal health status measured by their own vital signs.
What that means, that unlike teaching self-driving cars or identifying viruses or asteroids, humans are all different, therefore, it is not possible to generalise machine learning in healthcare to individuals. Meaning of course, that machine learning itself needs to be personalised but more technically relevant, we need to develop machine learning for N-of-1, that is precision personalisation.
Humans are not cars.
Think about that.

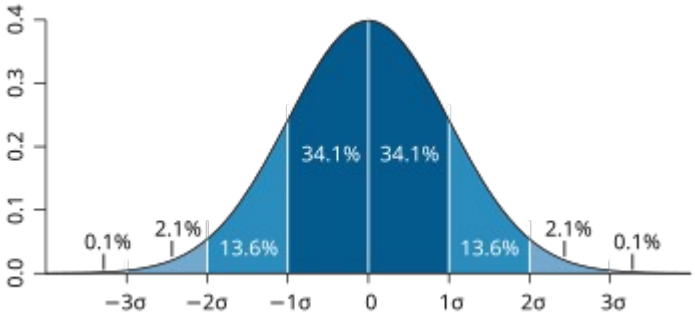

Figure 1: Humans or Cars?

Dr Mike Tremblay is a co-founder of AIS, a company focused on AI powered clinical decision support prediction machine for clinicians using real time vital signs, SDOH and environmental data.