Artificial Intelligence needs Clinical Intelligence to Succeed

John GF Cleland, MD #*

Charles Li, MSc #

Yola Jones, MSc #

# Robertson Centre for Biostatistics, Institute of Health and Wellbeing, University of Glasgow, Glasgow Royal Infirmary, Glasgow, UK.

* National Heart & Lung Institute, Imperial College, London.

Word Count for body of the text (excluding tables and footnotes) is 1,575

Address for Correspondence

Robertson Centre for Biostatistics & Glasgow Clinical Trials Unit,

University of Glasgow, UK. G12 8QQ

john.cleland@glasgow.ac.uk

"If I had asked people what they wanted, they would have said faster horses."
Attributed by some to Henry Ford.

"The answer to the ultimate question of life, the universe, and everything, calculated by an enormous supercomputer named DeepThought is 42. Unfortunately, no one knows what the question is." Douglas Adams in The Hitchhiker's Guide to the Galaxy.

In this issue of JACC, Jing et al investigate the application of machine learning (ML) to a large administrative set of health-care data for patients with heart failure (1). As most people with cardiovascular disease will develop heart failure before they die this is an issue of immense importance to health-care systems worldwide (2, 3). Importantly, Jing et al don't just try to build a better prognostic model but also to identify the effect of correcting deficiencies in care on mortality, overall and for each individual and each intervention. This approach has many benefits. It enables health-care providers to identify and prioritise individual patients whose management could be improved and to audit and enhance their overall organisational performance. It also provides an educational tool for health-care professionals and patients about which interventions will have the greatest effect on longevity and for whom. This takes us one step closer to delivering the right intervention, to the right patient at the right time; and avoiding the converse.

Patients with heart disease suffer from a surfeit of predictive models (4), which are usually based on patients selected for and willing to participate in clinical trials (5), with very little purpose other than making a better guess at which patients will die or be re-hospitalised. This academic output may have very little relevance to patient care, especially for patients with more unstable or more advanced heart failure for whom symptom-control rather than prognosis may be the key issue. For instance, only patients with severe symptoms that are not responding to guideline-recommended therapy will be selected for heart transplantation or a left ventricular assist device. Patients who respond to treatment will have such interventions deferred. Patients who fail to respond to therapy have a poor prognosis and there is no further practical need to apply risk scores. When creating predictive models, authors should be clear about the reasons, which is often given as identification of high-risk patients in need of closer monitoring and greater attention, although precisely what this entails is rarely specified and evidence that it makes a difference rarely provided. A cynic might say that the purpose of developing prognostic models is more to do with adding to the authors' list of publications than to improving scientific knowledge or patient-care.

What then should be the purpose of predictive models for heart failure? Ultimately, they should improve the efficacy, efficiency and quality of healthcare. They may indeed achieve this by identifying those who have an excellent prognosis with existing therapy who are unlikely to obtain benefit from further intervention or else by identifying those at high-risk who might be prioritised for further therapy, assuming that their risk can be favourably modified. Predictive models can also be used to audit outcomes within or amongst health-care systems. For instance, admissions for heart failure in the United Kingdom (UK) are associated with lengths of stay and mortality that are much higher than in the United States of America (US)(6). However, after risk adjustment, the mortality in the UK is exactly what models based on US data predict, suggesting that thresholds for hospital admission in the US are lower than for the UK, perhaps reflecting defensive medicine or financial incentives. A third useful purpose of predictive models is for education of patients, clinicians and organisations to show the benefits of an intervention, the harm done by their inadvertent omission and which interventions should be prioritised.

The concept of ML as a method of applying artificial intelligence to challenging problems has been around since the 1950s. There has been an exponential growth in applications of ML to many aspects of life over the last decade driven by a growth in the availability of data and in computing power. Medicine has lagged behind for several reasons including lack of financial incentives, lack of interest or scepticism from clinicians, lack of easy access to large, comprehensive, well-curated data-sets due to data-privacy regulations failure and, last but not least, lack of clinically intelligent questions. Recent reports of ML applications in medicine make it clear that many data-scientists don't understand the clinical issues and that few clinicians are sufficiently engaged with or provide constructive, critical evaluation. ML is an unbiased technique that only does what it is programmed to do, which may well be to provide an answer such as "42", as in the Hitchhiker's Guide to the Galaxy, and leave everyone wondering how to interpret the result.

Health-care administrative systems now provide access to large amounts of data, including the full range of patients seeking assistance and many variables that are not conventionally included in prognostic models. However, many patients may not be referred for specialist investigation or care and so the diagnosis of interest may not be robust. Essential data may be missing (eg:- echocardiographic results) and this will often not be at random, rendering imputation of missing-data potentially hazardous. Clinical trials and registries usually identify patients with a more robust diagnosis and provide more complete data but they exclude patients who do not come in contact with the specialists conducting the research, those who don't quite fit the selection criteria (often the largest and most interesting group of patients) and those who refuse to participate (Table). Well-curated data-sets should be of great value at a health service or population level for prognostic modelling but, for heart failure, most such administrative data-sets are sufficiently incomplete to be of uncertain value.

Jing et al applied ML to more than 200 variables collected in electronic health records from thirteen regional hospitals and a network of primary and specialty clinics belonging to the Geisinger health system(1, 7). The models were trained on a data-set of 26,971 patients with heart failure and either a reduced (HFrEF) or preserved (HFpEF) left ventricular ejection fraction. Of the ML models employed, one called XGBoost performed best, modestly improving the area-under-the sensitivity/specificity curve (AUC) for one-year mortality from 0.74 to 0.77. This is no better than that reported for many prognostic models using more conventional statistical techniques (8, 9). Possible reasons for the disappointing performance was the large amount of missing data that was imputed (48 variables had values for <50% of cases) and the age of the population (25% were aged ≥84 years); a key but unmodifiable driver of prognosis. Efficient health-care systems should focus on treating modifiable risk and compassionate management of that which is not.

The next step in the analysis was much more innovative. Three omissions of care (care-gaps) were considered generic to all heart failure phenotypes (percentages in brackets refer to patients who were eligible for an intervention but with a care-gap), including flu vaccination (59%), a haemoglobin A1c <8% (36% of diabetics) and blood pressure of >130/80mmHg (30%), although none of these care-gaps has yet been shown in randomised trials of heart failure to be actionable. Five care-gaps were relevant only to patients with HFrEF, including use of angiotensin converting enzyme inhibitors (ACEi) or angiotensin-II receptor blockers (ARB) with or without neprilysin inhibition (41%), mineralocorticoid receptor antagonists (74% excluding patients with severe renal dysfunction or hyperkalaemia), evidence-based beta-blockers (25%), hydralazine and nitrate (91% of those for whom ACEi/ARB were contraindicated) and cardiac resynchronisation therapy (65% of those with left bundle branch block and a QRS duration ≥150ms). At least four of these would be considered essential for good care for HFrEF. Clearly, care-gaps are common even in an excellent health-care system and audit can potentially have a large impact on performance. The model predicted that, of 13,238 patients alive in November 2019, 2,844 (21%) would die in the following year and that 231 deaths (only 8% fewer deaths) would be prevented by plugging care-gaps. Surprisingly, a similar number of deaths were predicted to be prevented for HFrEF and HFpEF, suggesting a substantial benefit from plugging the generic but poorly substantiated care-gaps. Importantly, much of this benefit was driven by small benefits in 8,897 lower-risk patients. Only 808 patients, just 6% of those at-risk, were predicted to be both at high-risk (~50% one-year mortality) and to receive a large benefit (absolute risk reduction ~18% or ~120 lives-saved) if the care-gaps were plugged. A much larger group of patients (3,452) had a predicted one-year mortality of 50% but were not expected to receive any benefit from plugging the care-gaps. These anomalies may be driven largely by age. Younger patients may be at lower risk but have greater modifiable risk; cardiovascular interventions may have little effect on risk in older patients. We should not presume that models based on the whole population work equally well for all sub-groups.

From a population perspective, plugging care-gaps that make a small difference to many lower-risk patients may save the greatest number of lives. However, from the perspectives of patients, clinicians, payers and health-care organisations, it makes more sense to focus on those for whom intervention makes the largest difference. Indeed, for heart failure, perhaps the greatest current unmet need is for the identification of people who have little to gain from further treatment, either because they are too well or too sick (10).

Ultimately, machine learning is just a method of searching for associations and is only as intelligent as the question being asked. The reasons for associations, especially when there is a lot of data missing for a reason, can be obscure and difficult to interpret. If you ask a dumb question, you should expect a dumb answer. There is nothing artificial about intelligence.

Footnote: Incidentally or by design "42" doesn't seem to be such a dumb answer – check it out on Wikipedia.

Sources of Information for Prognostic Models in Heart Failure

	Clinical Trials	Registries	Administrative Data
Usual Size	<5,000	<20,000	>100,000
Patients	Highly Selected	Selected	Relatively Unselected
Co-morbidity	Often Low	Intermediate	High
Risk Profile	Intermediate (high & low risk often excluded)	Variable	All Included
Older People	Under-represented	Under-represented	Highly Represented
Minority Groups	Under-represented	Variable	Reflects Epidemiology
Diagnosis of Heart Failure			
Diagnostic Specificity	Very High	High	Low
Diagnostic Sensitivity	Not addressed	Low	High
Symptoms & Signs	Collected	Collected	Variable / NLP*
LVEF: Bias	High	Intermediate	Low
LVEF: Quality	High	Intermediate	Variable
Atrial Volume	Variable	Variable	Variable
Natriuretic Peptides	If pre-specified	Variable	Variable
Treatment at Visits	Yes	Yes	Variable
All Treatment	Variable	Variable	Variable
Prognostic Variables			
Pre-specified	Yes	Yes	No
Number	Small	Small	Large
Frequency of Measurement	Low	Low	May be high
Missing if pre-specified	Low	Low	N/A
Missing if not pre-specified	Usually	Usually	Often
Well-Curated	Usually	Often	Often Not

Footnote

Clinical trials generally have well-curated data-sets, with a limited selection of variables, relatively few missing data on relatively few, highly-selected patients. There will be many biases for patient selection including those imposed by the protocol, those applied by the investigator (including biased interpretation of the protocol and those that can be attributed to patients who agree to participate. Clinical registries are similar to clinical trials in many respects, especially if they require patient-consent, but may be larger, enrol patients more representative of clinical practice, at least for the specialist group enrolling them, and may collect specific data that is not available in either trial or administrative data-sets. NLP* = natural language processing that enables information in reports and discharge/clinic letters to be extracted and added to data-sets.

Reference List

1. Jing L, et al. A Machine Learning Approach to Management of Heart Failure Populations. *Journal of the American College of Cardiology* 2020.

2. Torabi A, Cleland JGF, Khan NK, Loh PH, Clark AL, Alamgir F, Caplin JL, Rigby AS, Goode K. The timing of development and subsequent clinical course of heart failure after a myocardial infarction. *Eur Heart J* 2008;29(7):859-870.

3. Torabi A, Rigby AS, Cleland JGF. Declining In-Hospital Mortality and Increasing Heart Failure Incidence in Elderly Patients with First Myocardial Infarction. *J Am Coll Cardiol* 2009;55(1):79-81.

4. Damen JA, Hooft L, Schuit E, Debray TP, Collins GS, Tzoulaki I, Lassale CM, Siontis GC, Chiocchia V, Roberts C, Schlussel MM, Gerry S, Black JA, Heus P, van der Schouw YT, Peelen LM, Moons KG. Prediction models for cardiovascular disease risk in the general population: systematic review. *BMJ* 2016;353:i2416.

5. Clark AL, Lammiman MJ, Goode K, Cleland JG. Is taking part in clinical trials good for your health? A cohort study. *Eur J Heart Fail* 2009;11(11):1078-1083.

6. Nagai T, Sundaram V, Shoaib A, Shiraishi Y, Kohsaka S, Rothnie KJ, Piper S, McDonagh TA, Hardman SMC, Goda A, Mizuno A, Sawano M, Rigby AS, Quint JK, Yoshikawa T, Clark AL, Anzai T, Cleland JGF. Validation of U.S. mortality prediction models for hospitalized heart failure in the United Kingdom and Japan. *Eur J Heart Fail* 2018;20(8):1179-1190.

7. Wehner GJ, Jing L, Haggerty CM, Suever JD, Leader JB, Hartzel DN, Kirchner HL, Manus JNA, James N, Ayar Z, Gladding P, Good CW, Cleland JGF, Fornwalt BK. Routinely reported ejection fraction and mortality in clinical practice: where does the nadir of risk lie? *Eur Heart J* 2019.

8. Nagai T, Sundaram V, Rothnie K, Quint JK, Shoaib A, Shiraishi Y, Kohsaka S, Piper S, McDonagh TA, Hardman SMC, Goda A, Mizuno A, Kohno T, Rigby AS, Yoshikawa T, Clark AL, Anzai T, Cleland JGF. Mortality after admission for heart failure in the UK compared with Japan. *Open Heart* 2018;5(2):e000811.

9. Rahimi K, Bennett D, Conrad N, Williams TM, Basu J, Dwight J, Woodward M, Patel A, McMurray J, MacMahon S. Risk prediction in patients with heart failure: a systematic review and analysis. *JACC Heart Fail* 2014;2(5):440-446.

10. Cleland JGF, Tavazzi L, Daubert JC, Tageldien A, Freemantle N. Cardiac resynchronization therapy: are modern myths preventing appropriate use? *J Am Coll Cardiol* 2009;53(7):608-611.