OPEN ACCESS

Artificial intelligence, bias and clinical safety

Robert Challen,[1,2] Joshua Denny,[3] Martin Pitt,[4] Luke Gompels,[2] Tom Edwards,[2] Krasimira Tsaneva-Atanasova[1]

[1]EPSRC Centre for Predictive Modelling in Healthcare, University of Exeter College of Engineering Mathematics and Physical Sciences, Exeter, UK
[2]Taunton and Somerset NHS Foundation Trust, Taunton, UK
[3]Departments of Biomedical Informatics and Medicine, Vanderbilt University Medical Center, Nashville, Tennessee, USA
[4]NIHR CLAHRC for the South West Peninsula, St Luke's Campus, University of Exeter Medical School, Exeter, UK

Correspondence to
Dr Robert Challen, EPSRC Centre for Predictive Modelling in Healthcare, University of Exeter College of Engineering Mathematics and Physical Sciences, Exeter EX4 4QF, UK; rc538@exeter.ac.uk

Received 23 May 2018
Revised 23 November 2018
Accepted 6 December 2018
Published Online First 12 January 2019

Linked

► http://dx.doi.org/10.1136/bmjqs-2018-008551

Check for updates

To cite: Challen R, Denny J, Pitt M, *et al*. *BMJ Qual Saf* 2019;**28**:231–237.

INTRODUCTION

In medicine, artificial intelligence (AI) research is becoming increasingly focused on applying machine learning (ML) techniques to complex problems, and so allowing computers to make predictions from large amounts of patient data, by learning their own associations.[1] Estimates of the impact of AI on the wider economy globally vary wildly, with a recent report suggesting a 14% effect on global gross domestic product by 2030, half of which coming from productivity improvements.[2] These predictions create political appetite for the rapid development of the AI industry,[3] and healthcare is a priority area where this technology has yet to be exploited.[2 3] The digital health revolution described by Duggal *et al*[4] is already in full swing with the potential to 'disrupt' healthcare. Health AI research has demonstrated some impressive results,[5–10] but its clinical value has not yet been realised, hindered partly by a lack of a clear understanding of how to quantify benefit or ensure patient safety, and increasing concerns about the ethical and medico-legal impact.[11]

This analysis is written with the dual aim of helping clinical safety professionals to critically appraise current medical AI research from a quality and safety perspective, and supporting research and development in AI by highlighting some of the clinical safety questions that must be considered if medical application of these exciting technologies is to be successful.

TRENDS IN ML RESEARCH

Clinical decision support systems (DSS) are in widespread use in medicine and have had most impact providing guidance on the safe prescription of medicines,[12] guideline adherence, simple risk screening[13] or prognostic scoring.[14] These systems use predefined rules, which have predictable behaviour and are usually shown to reduce clinical error,[12] although sometimes inadvertently introduce safety issues themselves.[15 16] Rules-based systems have also been developed to address diagnostic uncertainty[17–19] but have struggled to deal with the breadth and variety of information involved in the typical diagnostic process, a problem for which ML systems are potentially better suited.

As a result of this gap, the bulk of research into medical applications of ML has focused on diagnostic decision support, often in a specific clinical domain such as radiology, using algorithms that learn to classify from training examples (supervised learning). Some of this research is beginning to be applied to clinical practice, and from these experiences lessons can be learnt about both quality and safety. Notable examples of this include the diagnosis of malignancy from photographs of skin lesions,[6] prediction of sight-threatening eye disease from optical coherence tomography (OCT) scans[7] and prediction of impending sepsis from a set of clinical observations and test results.[20 21]

Outside of diagnostic support ML systems are being developed to provide other kinds of decision support, such as providing risk predictions (eg, for sepsis[20]) based on a multitude of complex factors, or tailoring specific types of therapy to individuals. Systems are now entering clinical practice that can analyse CT scans of a patient with cancer and by combining this data with learning from previous patients, provide a radiation treatment recommendation, tailored to that patient which aims to minimise damage to nearby organs.[22]

Other earlier stage research in this area uses algorithms that learn strategies to maximise a 'reward' (reinforcement learning). These have been used to test

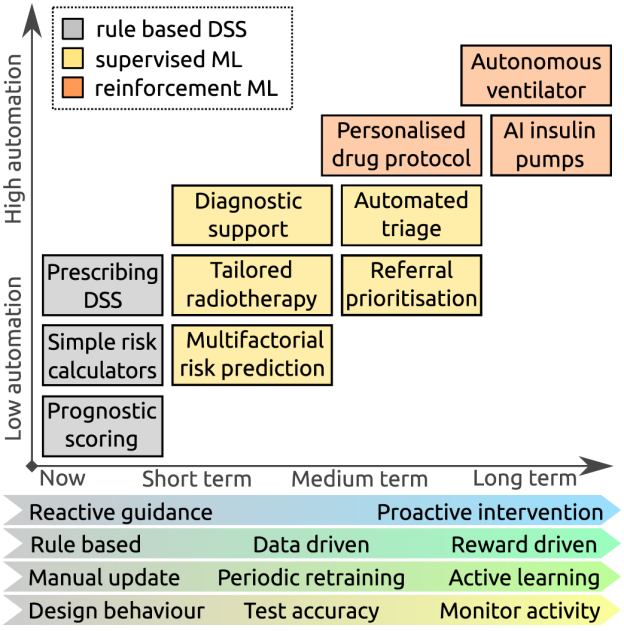

Figure 1 Expected trends in machine learning (ML) research: boxes show representative examples of decision support tasks that are currently offered by rule-based systems (grey), and hypothetical applications of ML systems in the future (yellow and orange), demonstrating increasing automation. The characteristics of the ML systems that support these tasks are anticipated to evolve, with systems becoming more proactive and reward driven, continuously learning to meet more complex applications, but potentially requiring more monitoring to ensure they are working as expected. AI, artificial intelligence; DSS, decision support systems.

approaches to other personalised treatment problems such as optimising a heparin loading regime to maximise time spent within the therapeutic range[23] or targeting blood glucose control in septic patients to minimise mortality.[24]

Looking further ahead AI systems may develop that go beyond recommendation of clinical action. Such systems may, for example, autonomously triage patients or prioritise individual's access to clinical services by screening referrals. Such systems could entail significant ethical issues by perpetuating inequality,[25] analogous to those seen in the automation of job applicant screening,[26] of which it is said that 'blind confidence in automated e-recruitment systems could have a high societal cost, jeopardizing the right of individuals to equal opportunities in the job market'. This is a complex discussion and beyond the remit of this article.

Outside of medicine, the cutting edge of AI research is focused on systems that behave autonomously and continuously evolve strategies to achieve their goal (active learning), for example, mastering the game of Go,[27] trading in financial markets,[28] controlling data centre cooling systems[29] or autonomous driving.[30 31] The safety issues of such actively learning autonomous systems have been discussed theoretically by Amodei *et al*[32] and from this work we can identify potential issues in medical applications. Autonomous systems are long way off practical implementation in medicine, but one can imagine a future where 'closed loop' applications, such as subcutaneous insulin pumps driven by information from wearable sensors,[33] or automated ventilator control driven by physiological monitoring data in intensive care,[34] are directly controlled by AI algorithms.

These various applications of ML require different algorithms, of which there are a great many. Their performance is often very dependent on the precise composition of their training data and other parameters selected during training. Even controlling for these factors some algorithms will not produce identical decisions when trained in identical circumstances. This makes it difficult to reproduce research findings and will make it difficult to implement 'off the shelf' ML systems. It is notable in ML literature that there is not yet an agreed way to report findings or even compare the accuracy of ML systems.[35 36]

Figure 1 summarises expected trends in ML research in medicine, over the short, medium and longer terms, with the focus evolving from reactive systems, trained to classify patients from gold standard cases, with a measurable degree of accuracy, to proactive autonomous systems which continuously learn from experience, whose performance is judged on outcome. Translation of ML research into clinical practice requires a robust demonstration that the systems function safely, and with this evolution different quality and safety issues present themselves.

QUALITY AND SAFETY IN ML SYSTEMS

In an early AI experiment, the US army used ML to try to distinguish between images of armoured vehicles hidden in trees versus empty forests.[1] After initial success on one set of images, the system performed no better than chance on a second set. It was subsequently found that the positive training images had all been taken on a sunny day, whereas it had been cloudy in the control photographs—the machine had learnt to discriminate between images of sunny and cloudy days, rather than to find the vehicles. This is an example of an unwittingly introduced bias in the training set. The subsequent application of the resulting system to unbiased cases is one cause of a phenomenon called 'distributional shift'.

Short-term issues

Distributional shift

Distributional shift[32] is familiar to many clinicians, who find previous experience inadequate for new situations, and have to operate, cautiously, outside of a 'comfort zone'. ML systems can be poor at recognising a relevant change in context or data, and this results in the system confidently continuing to make erroneous predictions based on 'out-of-sample' inputs.[32]

A mismatch between training and operational data can be inadvertently introduced, most commonly, as above, by deficiencies in the training data, but also by inappropriate application of a trained ML system to an unanticipated patient context. Such situations can be described as 'out-of-sample' input, and the need to cater for many such edge cases is described as the 'Frame problem'[25] of AI.

The limited availability of high quality data for training, correctly labelled with the outcome of interest, is a recurrent issue in ML studies. For example, when data are available it may have been collected as 'interesting cases' and not representative of the normal, leading to a sample selection bias.[6] In another example, the outcome may be poorly defined (eg, pneumonia) and variably assigned by experts, leading to a training set with poor reproducibility, and no 'ground truth' to learn associations.[9]

Inappropriate application of an ML system to a different context can be quite subtle. De Fauw *et al*[7] discovered their system worked well on scans from one OCT machine, but not another, necessitating a process to normalise the data coming from each machine, before a diagnostic prediction could be made. Similarly we anticipate that the system for diagnosing skin malignancy,[6] which was trained on pictures of lesions biopsied in a clinic, may not perform as well when applied to the task of screening the general population where the appearance of lesions, and patient's risk profile, is different.

In some cases, distributional shift is introduced deliberately. ML systems perform best when index cases and controls are approximately equal in the training set,[37] and this is not common in medicine. Imbalanced data sets may be 'rebalanced' by under-sampling or over-sampling, and without correction the resulting system will tend to over-diagnose the rare case.[38] Alternative approaches may 'boost' the significance of true positive or false negative cases depending on the application, which can lead, for example, to a model good for screening but poor for diagnosis.[39]

Over time disease patterns change, leading to a mismatch between training and operational data. The effect of this on ML models of acute kidney injury was studied by Davis *et al*,[40] who found that over time decreasing AKI incidence was associated with increasing false positives from their ML system, an example of prediction drift.

There are many different ML algorithms, and they perform differently under the challenge of distributional shift, and this 'may lead to arbitrary and sometimes deleterious effects that are costly to diagnose and address'.[41] It is notable however that the sepsis detection system mentioned above[20] has been successfully tested in the different context of a community hospital[5] despite being trained in intensive care, a potential distributional shift, and thus shows some capability of adaptation through 'transfer learning'.[38 42]

Insensitivity to impact

In the comparison between ML systems and expert dermatologists performed by Esteva *et al*,[6] both humans and machines find it difficult to discriminate between benign and malignant melanocytic lesions, but humans 'err on the side of caution' and over-diagnose malignancy. The same pattern was not observed for relatively benign conditions. While this decreases a clinician's apparent accuracy, this behaviour alteration in the face of a potentially serious outcome is critical for safety, and something that the ML system has to replicate. ML systems applied to clinical care should be trained not just with the end result (eg, malignant or benign), but also with the cost of both potential missed diagnoses (false negatives) and over-diagnosis (false positives).[43] During learning ML systems assess and maximise their performance based on a measure of accuracy obtained on predictions made from training data. Often this accuracy measure does not take into account real-world impacts, and as a result the ML system can be optimised for the wrong task, and comparisons to clinician's performance flawed.

Black box decision-making

One of the key differences between rule-based systems and the multitude of ML algorithms is the degree to which the resulting prediction can be explained in terms of its inputs. Some ML algorithms, particularly those based on artificial neural networks, make inscrutable predictions and for these algorithms it is harder to detect error or bias. This issue was demonstrated by the armoured vehicle detection system developed by the US army described above[1] and has been most studied in ML systems relying on image analysis.[6 9] To mitigate this, such systems can produce 'saliency maps' which identify the areas of, for example, the skin lesion[6] or the chest X-rays,[9] which most contributed to their prediction. However, outside of image analysis this inscrutability is harder to manage, and detection of bias in black box algorithms requires careful statistical analysis of the behaviour of the model in the face of changing inputs.[44 45]

Unsafe failure mode

The concept of confidence of prediction was mentioned in the context of distributional shift above. As with interpretability, not all ML algorithms produce estimates of confidence. If ML systems are opaque to interpretation, it becomes essential for the clinician to be aware whether the system believes its prediction is a sensible one. If the system's confidence is low, best practice design would be to fail-safe[46] and refuse to make a prediction either way.

Table 1 A general framework for considering clinical artificial intelligence (AI) quality and safety issues in medicine

Issue	Summary	Example
Short term		
Distributional shift	A mismatch between the data or environment the system is trained on and that used in operation, due to bias in the training set, change over time, or use of the system in a different population, may result in an erroneous 'out of sample' prediction.	The accuracy of a system which predicts impending acute kidney injury based on other health records data, became less accurate over time as disease patterns changed.[40]
Insensitivity to impact	A system makes predictions that fail to take into account the impact of false positive or false negative predictions within the clinical context of use.	An unsafe diagnostic system is trained to be maximally accurate by correctly diagnosing benign lesions at the expense of occasionally missing malignancy.[6]
Black box decision making	A system's predictions are not open to inspection or interpretation and can only be judged as correct based on the final outcome.	A X-Ray analysis AI system could be inaccurate in certain scenarios because of a problem with training data, but as a black box this is not possible to predict and will only become apparent after prolonged use.[9]
Unsafe failure mode	A system produces a prediction when it has no confidence in the prediction accuracy, or when it has insufficient information to make the prediction.	An unsafe AI decision support system may predict a low risk of a disease when some relevant data is missing. Without any information about the prediction confidence, a clinician may not realise how untrustworthy this prediction is.[46]
Medium term		
Automation complacency	A system's predictions are given more weight than they deserve as the system is seen as infallible or confirming initial assumptions.	The busy clinician ceases to consider alternatives when a usually predictable AI system agrees with their diagnosis.[48]
Reinforcement of outmoded practice	A system is trained on historical data which reinforces existing practice, and cannot adapt to new developments or sudden changes in policy	A drug is withdrawn due to safety concerns but the AI decision support system cannot adapt as it has no historical data on the alternative.
Self-fulfilling prediction	Implementation of a system indirectly reinforces the outcome it is designed to detect.	A system trained on outcome data, predicts that certain cancer patients have a poor prognosis. This results in them having palliative rather than curative treatment, reinforcing the learnt behaviour.
Long term		
Negative side effects	System learns to perform a narrow function that fails to take account of some wider context creating a dangerous unintended consequence.	An autonomous ventilator derives a ventilation strategy that successfully maintains short term oxygenation at the expense of long-term lung damage.[34]
Reward hacking	A proxy for the intended goal is used as a 'reward' and a continuously learning system finds an unexpected way to achieve the reward without fulfilling the intended goal.	An autonomous heparin infusion finds a way to control activated partial thromboplastin time (aPTT) at the time of measurement without achieving long-term control.[23]
Unsafe exploration	An actively learning system begins to learn new strategies by testing boundary conditions in an unsafe way.	A continuously learning autonomous heparin infusion starts using dangerously large bolus doses to achieve rapid aPTT control.[23]
Unscalable oversight	A system requires a degree of monitoring that becomes prohibitively time consuming to provide.	An autonomous subcutaneous insulin pump requires the patient to provide exhaustive detail of everything they have eaten before it can adjust the insulin regime.[33]

A similar fail-safe may be needed if the system has insufficient input information or detects an 'out-of-sample' situation as described above.[46]

Medium-term issues

Automation complacency

As humans, clinicians are susceptible to a range of cognitive biases which influence their ability to make accurate decisions.[47] Particularly relevant is 'confirmation bias' in which clinicians give excessive significance to evidence which supports their presumed diagnosis and ignore evidence which refutes it.[25] Automation bias[48] describes the phenomenon whereby clinicians accept the guidance of an automated system and cease searching for confirmatory evidence (eg, see Tsai *et al*[49]), perhaps transferring responsibility for decision-making onto the machine—an effect reportedly strongest when a machine advises that a case is normal.[48] Automation complacency is a related concept[48] in which people using imperfect DSS are least likely to catch errors if they are using a system which has been generally reliable, they are loaded with multiple concurrent tasks and they are at the end of their shift.

Automation complacency can occur for any type of decision support, but may be potentiated when combined with other pitfalls of ML described above. For example, given the sensitivity to distributional shift described, the usually reliable ML system that encounters an out-of-sample input may not 'fail safely' but continue confidently to make an erroneous prediction of low malignancy risk and not be

questioned by the busy clinician who then ceases to consider alternatives.

Reinforcement of outmoded practice and self-fulfilling predictions

In the medium term, we expect to see systems emerging from research that use ML to recommend the most appropriate clinical actions, for example, by identifying patients who might benefit most from a specific treatment or for whom further referral and investigation is warranted.[7]

Such recommendation decision support already exists, but in systems whose behaviour is determined by explicitly designed rules. The shift to a data-driven approach introduces a new risk in the situation of a sudden change in clinical practice that requires the DSS to change, for example, a drug safety alert. While the rule-based system can be manually updated, as ML is predicated on the availability of appropriate data, it has the potential to reinforce outmoded practice, and a radical change that invalidates historical practice is difficult to absorb, as there are no prior data to retrain the system with. The need to periodically retrain and evaluate performance in response to technological evolution, new knowledge and protocol changes in medicine requires costly updating of gold standard data sets.

On the other hand, a related potential problem could arise in ML systems that are very frequently updated, and particularly those that continuously learn. Suppose a system predicts a prognosis, this may in turn influence therapy in a way that reinforces the prognosis and lead to a positive feedback loop. In this scenario, there is a self-fulfilling prediction, which then may be further reinforced as the ML system learns.

Longer-term issues

Table 1 incorporates Amodei *et al*'s framework for safety in AI,[32] which deals with issues more specific to continuously learning, autonomous systems. For obvious reasons, such systems will be challenging to deploy in the context of medicine and so their safety issues are less immediate. Rather than repeating Amodei *et al*'s detailed analysis,[32] we describe these issues using hypothetical scenarios based on the research into personalised heparin dosing mentioned above[23]:

- Negative side effects: The target of maximising the time in the therapeutic window requires careful management of heparin infusions that delay administration of other medications
- Reward hacking: An automated system may find ways in which to 'game' the goals defined by the reward function. The heparin dosing system, for example, might stumble on a strategy of giving pulses of heparin, immediately before activated partial thromboplastin time (aPTT) measurement, giving good short-term control, but without achieving the intended goal of stable long-term control. This is known as 'hacking the reward function' or 'wireheading'.[32]
- Unsafe exploration: As part of its continuous learning, the system may experiment with the dosing of heparin to try and improve its current behaviour. How do we set limits to prevent dangerous overdosing, and define what changes in strategy are safe for the system to 'explore'[50]?
- Unscalable oversight: As the system is learning new strategies for heparin management for novel patient groups, the management strategies it proposes require inconveniently frequent and expensive aPTT measurement.

At present these issues are merely theoretical in medicine, but they have been observed in ML test environments[51] and are increasingly becoming relevant in applications such as autonomous driving systems.[31]

Box 1 - Quality control questions for short-term and medium-term issues in machine learning

Distributional shift
- Has the system been tested in diverse locations, underlying software architectures (such as electronic health records), and populations?
- How can we be sure the training data matches what we expect to see in real life and does not contain bias?
  - How can we be confident of the quality of the 'labels' the system is trained on?
  - Do the 'labels' represent a concrete outcome ('ground truth') or a clinical opinion?
  - How has imbalance in the training set been addressed?
  - Is the system applied to the same diagnostic context that it was trained in?
- How is the system going to be monitored and maintained over time to adjust for prediction drift?

Insensitivity to impact
- Does the system adjust its behaviour ('err on the side of caution') where there are high impact negative outcomes?
- Can the system identify 'out of sample' input and adjust its confidence accordingly?

Black box decision-making, unsafe failure and automation complacency
- Are the system's predictions interpretable?
- Does it produce an estimate of confidence?
- How is the certainty of prediction communicated to clinicians to avoid automation bias?

Reinforcement of outmoded practice and self-fulfilling predictions
- How can it accommodate breaking changes to clinical practice?
- What aspects of existing clinical practice does this system reinforce?

CONCLUSION

Developing AI in health through the application of ML is a fertile area of research, but the rapid pace of change, diversity of different techniques and multiplicity of tuning parameters make it difficult to get a clear picture of how accurate these systems might be in clinical practice or how reproducible they are in different clinical contexts. This is compounded by a lack of consensus about how ML studies should report potential bias, for which the authors believe the Standards for Reporting of Diagnostic Accuracy initiative[52] could be a useful starting point. Researchers need also to consider how ML models, like scientific data sets, can be licensed and distributed to facilitate reproduction of research results in different settings.

As ML matures we suggest a set of short-term and medium-term clinical safety issues (see table 1) that need addressing to bring these systems from laboratory to bedside. This framework is supported by a set of quality control questions (Box 1) that are designed to help clinical safety professionals and those involved in developing ML systems to identify areas of concern. Detailed mitigation of these issues is a large topic that cannot be addressed here, but is discussed by Amodei *et al*[32] and Varshney *et al*.[46]

Implementation of ML DSS in the short term is likely to focus on diagnostic decision support. ML diagnostic decision support should be assessed in the same manner and with the same rigour as the development of a new laboratory screening test. Wherever possible a direct comparison should be sought to existing decision support or risk scoring systems—ideally through a randomised controlled trial as exemplified by Shimabukuro *et al*.[42 53]

As with all clinical safety discussions we need to maintain a realistic perspective. Suboptimal decision-making will happen with or without ML support, and we must balance the potential for improvement against the risk of negative outcomes.

Acknowledgements The authors thank David Chalkley, Deputy CCIO & IT Clinical Safety Lead, TSFT, for comments that greatly enhanced this article.

Contributors All authors discussed the concept of the article and RC wrote the initial draft. KTA, JD, TE, MP and LG commented and made revisions, DC critically reviewed the draft. All authors agreed with the final manuscript. RC is the guarantor.

Funding This article was funded by Engineering and Physical Sciences Research Council and the grant number is EP/N014391/1.

Competing interests None declared.

Patient consent for publication Not required.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES

1 Dreyfus HL, Dreyfus SE. What artificial experts can and cannot do. *AI Soc* 1992;6:18–26.
2 Rao A, Verweij G, Cameron E. Sizing the prize: what's the real value of AI for your business and how can you capitalise? PwC. 2017. Available: https://www.pwc.com/gx/en/issues/analytics/assets/pwc-ai-analysis-sizing-the-prize-report.pdf
3 Hall W, Pesenti J. Growing the artificial intelligence industry in the UK - GOV.UK. Department for digital, culture, media & sport and department for business, energy & industrial strategy. 2017. Available: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/652097/Growing_the_artificial_intelligence_industry_in_the_UK.pdf
4 Duggal R, Brindle I, Bagenal J. Digital healthcare: regulating the revolution. *BMJ* 2018;360:k6.
5 McCoy A, Das R. Reducing patient mortality, length of stay and readmissions through machine learning-based sepsis prediction in the emergency department, intensive care unit and hospital floor units. *BMJ Open Qual* 2017;6:e000158.
6 Esteva A, Kuprel B, Novoa RA, *et al*. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 2017;542:115–8.
7 De Fauw J, Ledsam JR, Romera-Paredes B, *et al*. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nat Med* 2018;24:1342–50.
8 Walsh CG, Ribeiro JD, Franklin JC. Predicting risk of suicide attempts over time through machine learning. *Clin Psychol Sci* 2017:1–13.
9 Rajpurkar P, Irvin J, Zhu K. CheXNet: radiologist-level pneumonia detection on chest x-rays with deep learning. arXiv [cs.CV]. 2017. Available: http://arxiv.org/abs/1711.05225
10 Gulshan V, Peng L, Coram M, *et al*. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA* 2016;316:2402.
11 Char DS, Shah NH, Magnus D. Implementing machine learning in health care - addressing ethical challenges. *N Engl J Med* 2018;378:981–3.
12 Kaushal R, Shojania KG, Bates DW. Effects of computerized physician order entry and clinical decision support systems on medication safety: a systematic review. *Arch Intern Med* 2003;163:1409–16.
13 Hippisley-Cox J, Coupland C, Vinogradova Y, *et al*. Predicting cardiovascular risk in England and Wales: prospective derivation and validation of QRISK2. *BMJ* 2008;336:1475–82.
14 Bouch DC, Thompson JP. Severity scoring systems in the critically ill. *Cont Edu Anaesth Crit Care Pain* 2008;8:181–5.
15 Koppel R*et al*. Role of computerized physician order entry systems in facilitating medication errors. *JAMA* 2005;293:1197–203.
16 Han YY*et al*. Unexpected increased mortality after implementation of a commercially sold computerized physician order entry system. *Pediatrics* 2005;116:1506–12.
17 Miller RA. Medical diagnostic decision support systems--past, present, and future: a threaded bibliography and brief commentary. *J Am Med Inform Assoc* 1994;1:8–27.
18 Nurek M, Kostopoulou O, Delaney BC, *et al*. Reducing diagnostic errors in primary care. a systematic meta-review of computerized diagnostic decision support systems by the

LINNEAUS collaboration on patient safety in primary care. *Eur J Gen Pract* 2015;21(sup1):8–13.
19 Bond WF, Schwartz LM, Weaver KR, *et al*. Differential diagnosis generators: an evaluation of currently available computer programs. *J Gen Intern Med* 2012;27:213–9.
20 Calvert JS, Price DA, Chettipally UK, *et al*. A computational approach to early sepsis detection. *Comput Biol Med* 2016;74:69–73.
21 Desautels T, Calvert J, Hoffman J, *et al*. Prediction of sepsis in the intensive care unit with minimal electronic health record data: a machine learning approach. *JMIR Med Inform* 2016;4:e28.
22 Thompson RF, Valdes G, Fuller CD, *et al*. Artificial intelligence in radiation oncology: a specialty-wide disruptive transformation? *Radiother Oncol* 2018;129:421–6.
23 Ghassemi MM, Clifford GD. Optimal medication dosing from suboptimal clinical examples: A deep reinforcement learning approach 23. In:*2016 38th annual international conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2016: 2978–81.
24 Weng W-H, Gao M, He Z. Representation and reinforcement learning for personalized glycemic control in septic patients. arXiv [cs.LG]. 2017. Available: http://arxiv.org/abs/1712.00654
25 K-H Y, Kohane IS. Framing the challenges of artificial intelligence in medicine. *BMJ Qual Saf* 2019;28:238–41.
26 Faliagka E, Tsakalidis A, Tzimas G. An integrated e-recruitment system for automated personality mining and applicant ranking. *Internet Research* 2012;22:551–68.
27 Silver D, Huang A, Maddison CJ, *et al*. Mastering the game of Go with deep neural networks and tree search. *Nature* 2016;529:484–9.
28 Nuti G, Mirghaemi M, Treleaven P, *et al*. Algorithmic trading. *Computer* 2011;44:61–9.
29 Evans R (deepmind), Gao J (deepmind). DeepMind AI reduces google data centre cooling bill by 40%. Available: https://deepmind.com/blog/deepmind-ai-reduces-google-data-centre-cooling-bill-40/
30 Office of the Assistant Secretary for Research and Technology. Automated driving systems 2.0 A vision for safety. national highway traffic safety administration. 2017. Available: https://www.nhtsa.gov/document/automated-driving-systems-20-voluntary-guidance
31 IIHS Status Report newsletter. 2018. Available: https://www.iihs.org/externaldata/srdata/docs/sr5304.pdf
32 Amodei D, Olah C, Steinhardt J. *Concrete problems in AI safety. arXiv [cs.AI]*. 06565, 2016.
33 Bothe MK, Dickens L, Reichel K, *et al*. The use of reinforcement learning algorithms to meet the challenges of an artificial pancreas. *Expert Rev Med Devices* 2013;10:661–73.
34 Prasad N, Cheng L-F, Chivers C. A reinforcement learning approach to weaning of mechanical ventilation in intensive care units. arXiv [cs.AI]. 2017. Available: http://arxiv.org/abs/1704.06300
35 Forman G, Scholz M. Apples-to-apples in cross-validation studies: pitfalls in classifier performance measurement. *ACM SIGKDD Explorations Newsletter* Published Online First. 2010. Available: https://dl.acm.org/citation.cfm?id=1882479
36 Lobo JM, Jiménez-Valverde A, Real R. AUC: a misleading measure of the performance of predictive distribution models. *Glob Ecol Biogeogr* 2008;17:145–51.
37 Haixiang G, Yijing L, Shang J, *et al*. Learning from class-imbalanced data: review of methods and applications. *Expert Syst Appl* 2017;73:220–39.
38 Storkey AJ. When Training and Test Sets are Different: Characterising Learning Transfer. In: Lawrence CSS, ed. *Dataset shift in machine learning*. MIT Press, 2013: 3–28.
39 Bae S-H, Yoon K-J. Polyp detection via imbalanced learning and discriminative feature learning. *IEEE Trans Med Imaging* 2015;34:2379–93.
40 Davis SE, Lasko TA, Chen G, *et al*. Calibration drift in regression and machine learning models for acute kidney injury. *J Am Med Inform Assoc* 2017;24:1052–61.
41 Sculley D, Phillips T, Ebner D. Machine learning: the high-interest credit card of technical debt. 2018. Available: https://research.google.com/pubs/pub43146.htmlhttp://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.675.9675 [Accessed 5 Mar 2018].
42 Mao Q, Jay M, Hoffman JL, *et al*. Multicentre validation of a sepsis prediction algorithm using only vital sign data in the emergency department, general ward and ICU. *BMJ Open* 2018;8:e017833.
43 Megler V, Gregoire S. Training models with unequal economic error costs using Amazon sagemaker. AWS machine learning blog. 2018. Available: https://aws.amazon.com/blogs/machine-learning/training-models-with-unequal-economic-error-costs-using-amazon-sagemaker/ [Accessed 19 Oct 2018].
44 Adler P, Falk C, Friedler SA, *et al*. *Auditing black-box models for indirect influence. arXiv [stat.ML]*, 2016.
45 Caruana R, Lou Y, Gehrke J. Intelligible models for healthcare: predicting pneumonia risk and hospital 30-day readmission. In:*Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. NY, USA: ACM, 2015: 1721–30.
46 Varshney KR. Engineering safety in machine learning. In:*2016 Information Theory and Applications Workshop*. ITA, 2016: 1–5.
47 Dawson NV, Arkes HR. Systematic errors in medical decision making: judgment limitations. *J Gen Intern Med* 1987;2:183–7.
48 Parasuraman R, Manzey DH. Complacency and bias in human use of automation: an attentional Integration. *Hum Factors* 2010;52:381–410.
49 Tsai TL, Fridsma DB, Gatti G. Computer decision support as a source of interpretation error: the case of electrocardiograms. *J Am Med Inform Assoc* 2003;10:478–83.
50 Garcia J, Fernandez F, Fern F. Safe exploration of state and action spaces in reinforcement learning. *J Artif Intell Res* 2012;45:515–64.
51 Leike J, Martic M, Krakovna V. AI safety gridworlds. arXiv [cs.LG]. 2017. Available: http://arxiv.org/abs/1711.09883
52 Bossuyt PM, Reitsma JB, Bruns DE, *et al*. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. Standards for Reporting of Diagnostic Accuracy. *Clin Chem* 2003;49:1–6.
53 Shimabukuro DW, Barton CW, Feldman MD, *et al*. Effect of a machine learning-based severe sepsis prediction algorithm on patient survival and hospital length of stay: a randomised clinical trial. *BMJ Open Respir Res* 2017;4:e000234.