Protecting patients from problematic predictions: the hidden risks of predictive clinical decision support tools.

Dr. Joseph Alderman[1*], MBChB, FRCA
Dr. Xiaoxuan Liu[2†], MBChB, PhD
Dr. Dhruv Parekh[3*], MBBS, PhD
Prof. Richard Riley[4†], PhD
Prof. Charlotte Summers[5‡], MBBS, PhD
Prof. Alastair Denniston[6*], MA, PhD

*1 Anesthesiology & critical care doctor, PhD student; 2 Ophthalmology doctor, clinician scientist; 3 Pulmonology & critical care doctor, associate professor; 4 Medical statistician, professor; 5 Pulmonology & critical care doctor, professor; 6 Ophthalmology doctor, professor | * University of Birmingham & University Hospitals Birmingham NHS Foundation Trust; † University of Birmingham; ‡ University of Cambridge & Cambridge University Hospitals NHS Foundation Trust.*

1: Background & Methods

PCDSTs (predictive clinical decision support tools) have been used for decades to support clinical decisions.[1,2] They are available for a wide variety of healthcare tasks, including assisting health professionals make diagnoses, assess patients' prognosis, and estimate risk of complications.

They include:

1. Simple clinical scoring systems, eg. CURB65 to classify the severity of community acquired pneumonia: Confusion, blood Urea nitrogen, Respiratory rate, Blood pressure, Age > 65).[3]
2. More complex predictive / statistical models, eg. This excerpt from the NELA risk score formula:

   Logit(predicted risk) = (-3.04678 + 0.06660 x Age_cent) + (1.13007 x ASA[3]) - (0.04323 x Albumin) + (0.01265 x Pulse_cent) **...** + (0.29453 x Soiling[Free bowel content, pus or blood]).[4]

3. AIaMD (Artificial Intelligence as a Medical Device), a subset of SaMD (Software as a Medical Device) which is regulated by the US FDA and other international medical device regulators.

Many of these tools are embedded directly within electronic health record systems (EHRs). Others are accessed via websites, apps, spreadsheet macros or other digital interfaces.

Many key clinical decisions can be influenced by PCDSTs, including whether to commence medications or admit patients to hospital, to determine if surgery or referral to intensive care are appropriate, and when planning discharge (**Fig 1**).

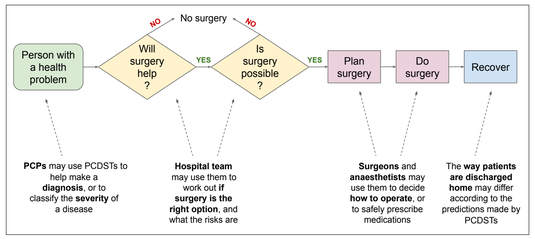

Figure 1: *Illustration of how PCDSTs might influence the care patients receive, using the perioperative care pathway as an example (the care of patients awaiting, receiving, or recovering from surgery).*

PCDSTs are frequently recommended in clinical guidelines to provide **quantified risk estimates**. For example, the UK's National Institute of Health and Care Excellence (NICE) guideline NG185[5] states:

"As soon as the diagnosis of unstable angina or NSTEMI is made[...] assess individual risk of future adverse cardiovascular events using an established risk scoring system that predicts 6-month mortality."

However, PCDTSs may give inaccurate predictions which mislead patients and clinicians when making treatment decisions.

- Methodological shortcomings,[6,7] inconsistent reporting, limited transparency,[8,9] and risk of bias [10] reduce PCDSTs' clinical utility.
- Unlike AIaMD, clinical scoring systems and predictive / statistical models are not effectively scrutinised by medical device regulators.[11]

The effect is that patients may receive treatments which do not align with their values and wishes.

This research programme used a convergent parallel mixed-methods design to uncover 'gaps' in the regulation / oversight of PCDSTs:

1. A national survey of UK surgeons, anesthesiologists, geriatricians and critical care physicians.
2. A semi-structured interview study involving 23 experts in clinical practice, statistics, healthcare regulation, research funding and PCDST development.
3. A patient focus group exploring PCDSTs and clinical risk communication from the perspective of 10 members of the public.

2: National survey

- Overall, 87 different PCDSTs are used in UK perioperative practice. Extrapolated across other countries and medical specialities **it is likely that hundreds of different PCDSTs are used in patients' care**.
- Many had incompletely overlapping functions, eg. 29 predict the risk of mortality but at different timepoints (perioperative, inpatient, or at 28, 30 or 90 days).
- **Several widely-used PCDSTs were first published decades ago**. For example, the Charlson Comorbidity Index was developed using data from 1984, and evaluated using data from 1969. Historical associations between particular diagnoses and adverse outcomes may no longer hold true. Similarly, understanding of socially constructed phenomena such as 'gender' or 'ethnicity' are in constant flux. **PCDSTs which are not updated to account for these 'data shifts' may show worse performance over time**.
- Clinicians report that many of these tools have **high influence** over the care they provide (**Fig 2**), and that they choose which tools to use based on external guidance from professional bodies or colleagues rather than reviewing evidence themselves. (**Fig 3**).
- The most used PCDSTs were different to those cited in a recent systematic – the number of citations is not a good surrogate of a tool's use in real-world practice.

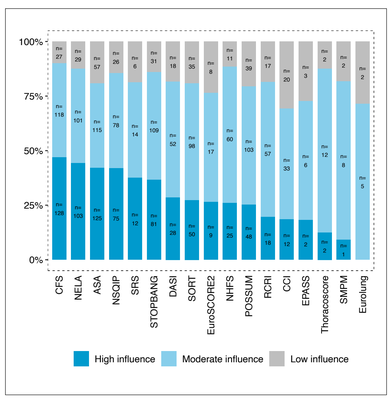

Figure 2: *The degree of influence the subset of the most widely used PCDSTs has on clinical judgement. Overall each tool had at least moderate influence for the majority of respondents who use them. The most influential tools were the Clinical Frailty Scale, the Parsimonious NELA risk calculator and ASA-PS.*

ASA = American Society of Anesthesiologists; CCI = Charlson Comorbidity Index; CFS = Clinical Frailty Scale; DASI = Duke Activity Status Index; EPASS = Estimation of Physiologic Ability and Surgical Stress; NELA = Parsimonious National Emergency Laparotomy Audit Risk calculator; NHFS = Nottingham Hip Fracture Score; NSQIP = National Surgical Quality Improvement Project surgical risk calculator; POSSUM = Physiologic and Operative Severity Score for the Study of Mortality and Morbidity; RCRI = Revised Cardiac Risk Index; SMPM = Surgical Mortality Prediction Model; SORT = Surgical Outcome Risk Tool; SRS = Surgical Risk Scale; STOPBANG = Snoring, Tired, Observed nocturnal apnoea, high blood Pressure, BMI, Age, Neck circumference, Gender.

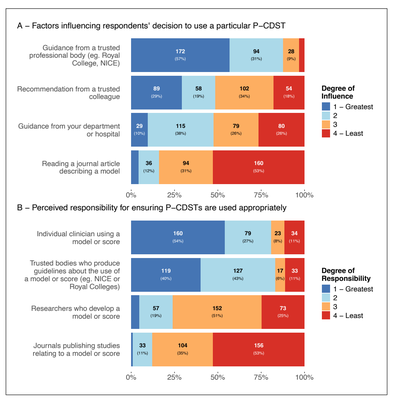

Figure 3: *The factors influencing respondents' decision to use a particular PCDSTs in their practice, and their perceptions regarding responsibility for ensuring that PCDSTs are used appropriately.*

3: Semi-structured interview study

Reflexive thematic analysis [12] **of the interview transcripts** using the lens of critical realism revealed that professionals believe that the current system of oversight for PCDSTs is inadequate (Theme 1). Participants perceived PCDSTs to be under-regulated, risking tools underperforming or demonstrating bias. They highlighted the need for greater guidance for clinicians, developers, and publishers of these tools (**Fig 4**):

"I would suspect that the clinical risk scores are probably very poorly regulated, if regulated at all, and I would suspect that it is very much a laissez-faire approach to them of just people saying this is a clinical risk score, use at your own peril." (Participant 003)

- They cautioned that whilst change is needed, greater harm could come from actions which risk continuity of care pathways dependent on PCDSTs (Theme 2).
- Participants also commented that regulatory frameworks (such as 'software as a medical device') were not designed with PCDSTs in mind, so may not be easy to apply for these tools (Theme 3).
- Improving the system of academic publishing was cited as a priority to improve the quality of studies developing and evaluating PCDSTs, in particular re-thinking the way academic 'credit' is allocated based on number of publications (Theme 4).

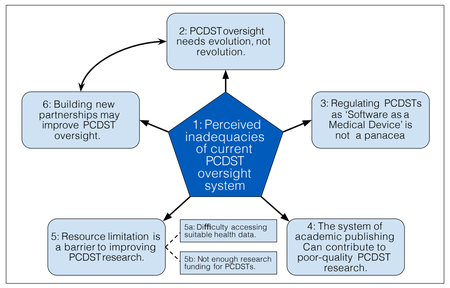

Figure 4: *Relationship between themes developed during reflexive thematic analysis of interview transcripts.*

- Many participants flagged the need to increase both research funding and access to high-quality data to enable high quality PCDSTs to be developed and evaluated (Themes 5a & 5b).
- Participants favoured taking an interdisciplinary approach when building a future system to improve PCDST assurance and oversight (Theme 6).

4: Focus group

Patients who participated in the focus group clearly stated that they did not know PCDSTs existed.

- They worried that their care could be biased or otherwise negatively influenced by these tools.
- They expressed a clear desire for transparency: they would want to know if and how PCDSTs are used by clinicians treating them.
- There was a strong desire to be treated as individuals, and scepticism about the concept of automated decision making (whether based on AI, PCDSTs or any other data-driven system).

5: Improving the governance of PCDSTs.

PCDSTs which give reliable predictions are helpful tools. However, many are subject to very limited scrutiny, and may never be re-evaluated once adopted into practice (Fig 5).

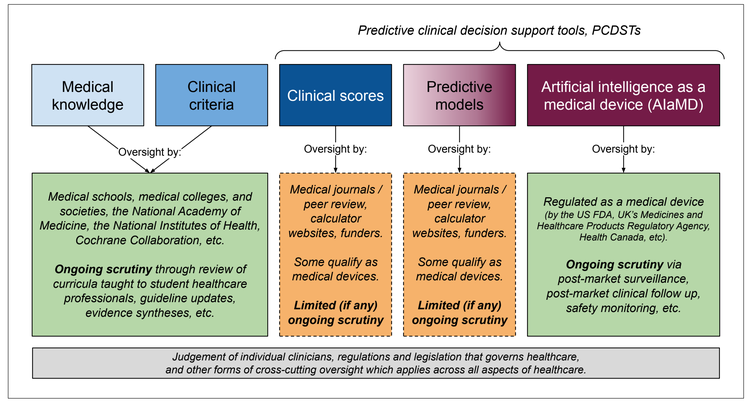

Figure 5: *The current structures providing oversight of PCDSTs. Whilst medical knowledge and clinical criteria are regularly re-evaluated by educational institutions and professional bodies, clinical scores and predictive models are subject to a single point evaluation during medical journal peer review. Once adopted into practice there is little scrutiny of these tools beyond intermittent and sporadic validation / evaluation studies. Moreover, there are few if any levers which can be used to restrict the use of problematic tools. In contrast, PCDSTs which qualify as AIaMD are subject to medical device regulation, which includes an obligation to monitor the ongoing performance and safety of the tool. In addition to regulation and oversight of individual technology types, broad horizontal oversight is provided in the form of the professional judgment of clinicians and their regulation as professionals, and by cross-cutting regulators.*

We advocate for a comprehensive review of the oversight and governance of PCDSTs:

Recommendation 1: The adequacy of existing oversight and goverance frameworks should be assessed, in particular the applicability of AIaMD / SaMD regulations to PCDSTs.

Recommendation 2: Robust guidance should be authored for clinicians who use PCDSTs in their practice, to enable them to better understand the tools they are using and their limitations.

Recommendation 3: Guidance about methodological best practice should be provided to developers of PCDSTs, enabling them to better understand the implication of design choices and datasets on the tools they create.

Recommendation 4: A mechanism should be established which enables the post-deployment monitoring of PCDSTs in real-world clinical use (post-market surveillance, or PMS in regulatory language).

We are building an interdisplinary research partnership to address these recommendations.

As a first step, we plan to build a platform promoting **transparency**, akin to those for registering clinical trials and systematic reviews. This will provide clarity on PCDSTs' purpose, supporting evidence, and other factors relevant when considering whether to use a particular tool.

- With sufficient resources, this may provide clinicians guidance about which PCDSTs would be appropriate to use in a particular clinical scenario (if any).
- To learn more, or to explore opportunities for collaboration please reach out via email at the address below.

Key takeaways:

1. *Predictive clinical decision support tools (PCDSTs) are widely used throughout healthcare, including within electronic health records (EHRs).*
2. *Whilst some are regulated as 'medical devices', many are subject to very limited scrutiny.*
3. *The governance and oversight of PCDSTs must be improved to provide assurance to clincians and patients. In particular: tool development, usage, monitoring, and regulation.*

References: *(See also project website - scan QR code in bottom right)*

1 – Challener et al, 2019 (10.1001/jama.2019.5284). 2 – van Smeden et al, 2021 (10.1016/j.jclinepi.2021.01.009). 3 – Lim et al, 2003 (10.1136/thorax.58.5.377).
4 – https://www.nela.org.uk/downloads/Technical%20Document%20NELA_PRS_Overview_Coefficients%20April%202023.pdf 5 – https://www.nice.org.uk/guidance/ng185 6 – Steyerberg et al, 2013 (10.1371/journal.pmed.1001381).
7 – Mallett et al, 2010 (10.1186/1741-7015-8-20). 8 – Collins et al, 2011 (10.1186/1741-7015-9-103). 9 – Collins et al, 2024 (10.1136/bmj-2023-078378). 10 – Vernooij et al, 2023 (10.1111/anae.15988). 11 – Sharma et al, 2021 (10.1136/bmjhci-2021-100421).
12 – Braun & Clarke. Thematic analysis: a practical guide. 2022 (ISBN: 978-1-4739-5324-6).

j.e.alderman@bham.ac.uk @jaldmn
www.linkedin.com/in/joseph-alderman

PeriopPrediction.co.uk