BMJ Digital Health & AI

Déjà vu in healthcare AI: lessons from the world's pioneer AI clinical decision support system

David Wong ,[1] Ruth Evans,[1] Niels Peek ,[2] Owen Johnson,[3] Susan Clamp[4]

To cite: Wong D, Evans R, Peek N, *et al*. Déjà vu in healthcare AI: lessons from the world's pioneer AI clinical decision support system. *BMJ Digit Health* 2026;**2**:e000226. doi:10.1136/bmjdhai-2025-000226

Received 10 September 2025
Accepted 25 November 2025

[1]Leeds Institute of Health Sciences, University of Leeds, Leeds, England, UK
[2]THIS Institute, University of Cambridge, Cambridge, UK
[3]School of Computer Science, University of Leeds, Leeds, UK
[4]Independent Researcher, Leeds, UK

Correspondence to
Dr David Wong, Leeds Institute of Health Sciences, University of Leeds, Leeds, UK;
d.c.wong@leeds.ac.uk

Recent advances in artificial intelligence (AI) have renewed interest in the possibility of computers assisting, or even replacing, doctors in making clinical decisions. However, computerised clinical decision support (CCDS) is not new, with scientific roots going back to the 1950s.[1] One of the first applications, a system for diagnosing causes of abdominal pain known as AAPHelp, was developed at the University of Leeds under Tim de Dombal's leadership.[2] To use the system, clinicians would take a structured assessment of a patient on a paper form (figure 1). Data from the form was then entered into a computer programme—a Naïve Bayes classifier, a simple machine learning algorithm that uses Bayes' rule to estimate conditional probabilities. The output was a differential diagnosis, in which each potential diagnosis had an estimated probability. In its initial installation, the programme ran on the 4.7 tonne KDF-9 computer and diagnoses could take up to 20 min. The aim of this editorial is to summarise the key findings from AAPHelp studies, contextualising them against the current AI zeitgeist and highlighting their continued relevance for today's AI research.

AAPHelp was implemented in multiple hospitals over two decades. During this time, its clinical impact was evaluated in a series of carefully designed studies. Many challenges that the Leeds team described are being rediscovered as modern researchers attempt to deploy new deep-learning CCDS. One such rediscovery is domain generalisability, in which AI models developed in one setting do not perform well when applied to another, superficially similar, setting. Lea and Jones recently highlighted how this problem was observed when AAPHelp's accuracy dropped significantly when tested at a new site in Copenhagen.[3 4] This was addressed in later versions of AAPHelp by using larger and more diverse training data from multiple countries. Later international studies, conducted with >15 000 patients, showed much smaller differences in accuracy between sites.[5]

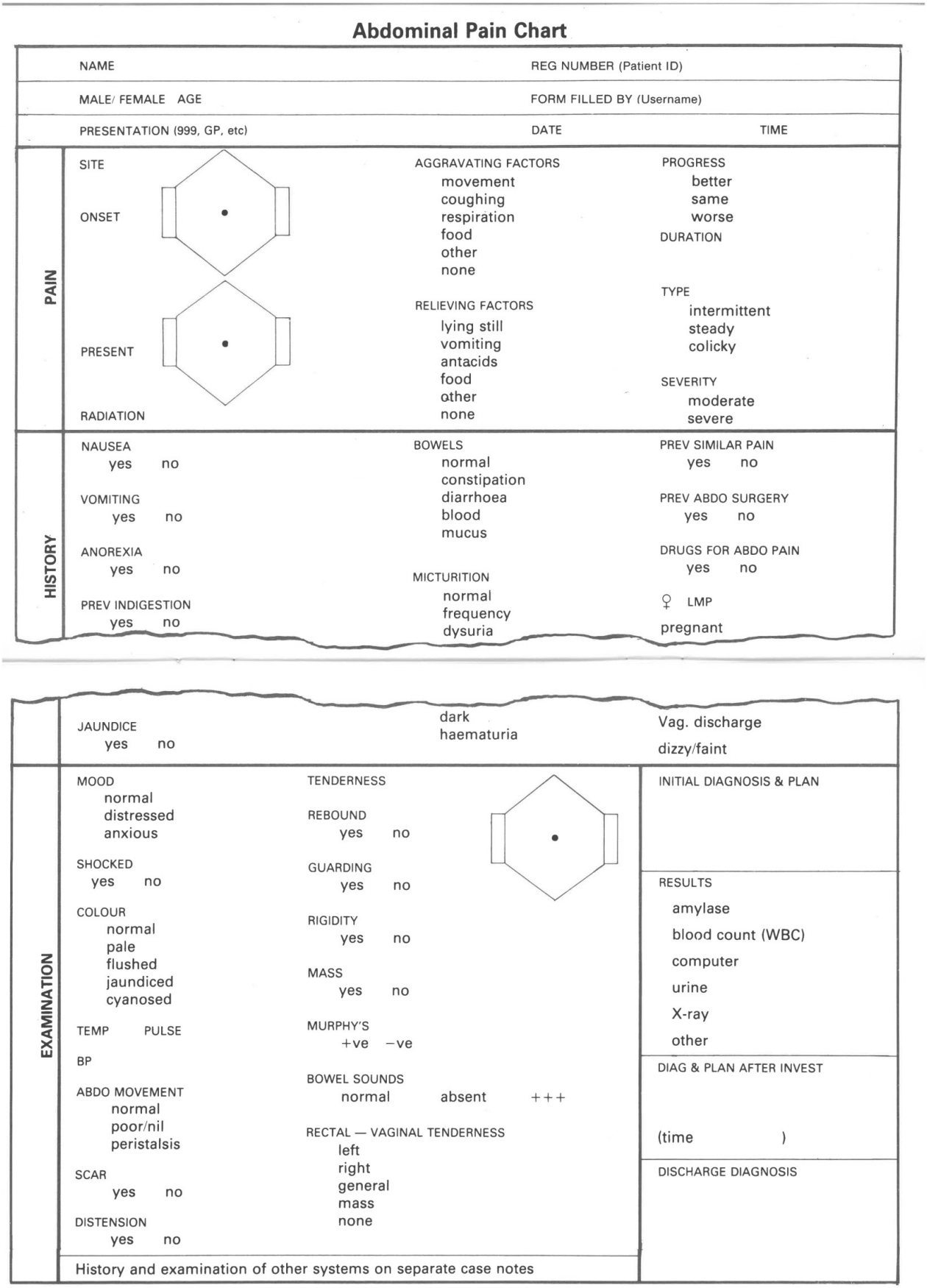
AI SYSTEMS REQUIRE CAREFUL CLINICAL EVALUATION

AAPHelp was clinically evaluated in eight UK hospitals in one of the first randomised trials of CCDS.[6] The researchers had the awareness that a straightforward 'CCDS' versus 'no CCDS' comparison (still the most prevalent approach to clinical CCDS evaluation) would not provide the desired insights. Instead, they used a factorial design to assess the individual impact of structured data collection, CCDS outputs and clinician performance feedback. Results showed that, after 6 months, the use of structured forms improved diagnostic accuracy against previous standard care from 45.7% to 56.7%. Adding feedback on patient outcomes improved accuracy further, approaching that of sites using the full system. The group concluded that reasons for improvement were multifaceted, but that changes in clinical process including discipline in data collection had greater impact on diagnostic accuracy than computer advice.

By modularising the CCDS intervention, Adams *et al* discerned whether the computer itself was responsible for clinical improvement. Such approaches are now well-established in medical informatics (see for example, Coiera[7]) but remain challenging to implement.

OUTCOMES IN CLINICAL EVALUATIONS

Another unique aspect of the multicentre evaluation was the choice of outcome metrics. In addition to reporting diagnostic accuracy and relevant clinical outcomes (eg, negative laparotomy rate), the study also reported outcomes related to the wider operation of



Abdominal Pain Chart

NAME
REG NUMBER (Patient ID)
MALE/ FEMALE AGE
FORM FILLED BY (Username)
PRESENTATION (999, GP, etc)
DATE
TIME

PAIN
SITE
ONSET
PRESENT
RADIATION
AGGRAVATING FACTORS
movement
coughing
respiration
food
other
none
RELIEVING FACTORS
lying still
vomiting
antacids
food
other
none
PROGRESS
better
same
worse
DURATION
TYPE
intermittent
steady
colicky
SEVERITY
moderate
severe

HISTORY
NAUSEA yes no
VOMITING yes no
ANOREXIA yes no
PREV INDIGESTION yes no
BOWELS
normal
constipation
diarrhoea
blood
mucus
MICTURITION
normal
frequency
dysuria
PREV SIMILAR PAIN yes no
PREV ABDO SURGERY yes no
DRUGS FOR ABDO PAIN yes no
♀ LMP
pregnant

JAUNDICE yes no
dark
haematuria
Vag. discharge
dizzy/faint

EXAMINATION
MOOD
normal
distressed
anxious
SHOCKED yes no
COLOUR
normal
pale
flushed
jaundiced
cyanosed
TEMP PULSE
BP
ABDO MOVEMENT
normal
poor/nil
peristalsis
SCAR yes no
DISTENSION yes no
TENDERNESS
REBOUND yes no
GUARDING yes no
RIGIDITY yes no
MASS yes no
MURPHY'S +ve −ve
BOWEL SOUNDS
normal absent +++
RECTAL — VAGINAL TENDERNESS
left
right
general
mass
none
INITIAL DIAGNOSIS & PLAN
RESULTS
amylase
blood count (WBC)
computer
urine
X-ray
other
DIAG & PLAN AFTER INVEST
(time)
DISCHARGE DIAGNOSIS
History and examination of other systems on separate case notes

Figure 1 An image of the structured paper form used to collect abdominal pain history for the AAPHelp computerised clinical decision support system. BP, blood pressure; GP, general practitioner; LMP, last menstrual period; WBC, white blood cell count.

the hospital, showing reductions in emergency department admissions and estimated cost savings.

The Leeds team recognised that demonstrating algorithmic accuracy is insufficient—its clinical impact needs to be evaluated across a range of clinically relevant outcomes. Impact depends not only on accuracy but crucially also on human and organisational factors. We note that, just as evaluation ought to be multifaceted, there is a growing call for wider outcomes to be integrated during AI training, ensuring alignment with intended use.[8]

LONG-TERM DATA DRIFT

One of the final evaluations of AAPHelp analysed diagnostic accuracy in a Scottish centre over a 15-year period.[9] AAPHelp initially outperformed all clinicians, but its accuracy declined from 78% to 55%, falling below the average for resident doctors, which remained stable. One of the original AAPhelp researchers (author SC) notes the unpublished finding that clinical users of the system tended to stop using it for more 'obvious' cases over time, leaving the algorithm to only provide predictions for more complex cases.

This presents one of the first known examples of longitudinal data drift, in which CCDS performance degrades as data (either observation or outcome) change over time. Modern AI algorithms are equally vulnerable, and the typical solution—regular updates—introduces further risks, including regulatory compliance. The UK's Medical and Healthcare products Regulatory Agency recently addressed this via an expert working group report.[10]

A CALL FOR CAREFUL EVALUATION

The history of AAPHelp underscores that many challenges in evaluating and implementing AI in healthcare are longstanding issues. Revisiting these early experiences offers valuable insights into pitfalls, such as generalisability, metric selection and data drift—lessons that remain relevant in the deep learning era. As AI rekindles enthusiasm for clinical decision support, its success will ultimately depend not on technological novelty alone, but on careful, rigorous and context-sensitive evaluation.

Contributors DW conceived of, and drafted the manuscript. DW, NP and RE developed the core themes. SC provided additional first-hand knowledge of computerised decision support in Leeds. All authors edited and reviewed the manuscript.

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests David Wong is an Associate Editor for *BMJ Digital Health & AI* but had no part in processing or review. The other authors declare no competing interests.

Patient and public involvement Patients and/or the public were not involved in the design, or conduct, or reporting, or dissemination plans of this research.

Patient consent for publication Not applicable.

Ethics approval Not applicable.

Provenance and peer review Commissioned; externally peer reviewed.

Data availability statement Data sharing not applicable as no datasets generated and/or analysed for this study.

ORCID iDs
David Wong https://orcid.org/0000-0001-8117-9193
Niels Peek https://orcid.org/0000-0002-6393-9969

REFERENCES

1 Ledley RS, Lusted LB. Reasoning Foundations of Medical Diagnosis. *Science* 1959;130:9–21.
2 Horrocks JC, McCann AP, Staniland JR, *et al*. Computer-aided Diagnosis: Description of an Adaptable System, and Operational Experience with 2,034 Cases. *BMJ* 1972;2:5–9.
3 Lea AS, Jones DS. Mind the Gap - Machine Learning, Dataset Shift, and History in the Age of Clinical Algorithms. *N Engl J Med* 2024;390:293–5.
4 Bjerregaard B, Brynitz S, Holst-Christensen J, *et al*. Computer-aided diagnosis of the acute abdomen: a system from leeds used on copenhagen patients. In: de Dombal FT, Gremy F, eds. *Decision making and medical care: can information science help?*. Amsterdam: North-Holland, 1976.
5 De Dombal FT, De Baere H, Van Elk PJ, *et al*. Objective medical decision making: acute abdominal pain. In: Beneken JEW, Thevenin V, eds. *Advances in biomedical engineering: results of the 4th EC Medical and Health Research Programme. Vol. 7 of Studies in health technology and informatics*. Burke, Va: IOS Press, 1993: 65–87.
6 Adams ID, Chan M, Clifford PC, *et al*. Computer aided diagnosis of acute abdominal pain: a multicentre study. *Br Med J (Clin Res Ed)* 1986;293:800–4.
7 Coiera E. Assessing technology success and failure using information value chain theory. In: *Applied interdisciplinary theory in health informaticsx*. IOS Press, 2019: 35–48.
8 Reyna MA, Nsoesie EO, Clifford GD. Rethinking Algorithm Performance Metrics for Artificial Intelligence in Diagnostic Medicine. *JAMA* 2022;328:329–30.
9 Stonebridge PA, Freeland P, Rainey JB, *et al*. Audit of computer-aided diagnosis of abdominal pain in accident and emergency departments. *Arch Emerg Med* 1992;9:271–3.
10 Rotalinti Y, Ordish J, Liu X, *et al*. Identifying and handling significant change due to data drift when assessing AI models in healthcare. *BMJ Digital Health* 2026. (In press).