frontiers

Clinical Deployment Environments: The Five Pillars of Translational Machine Learning for Health

Steve Harris [1,2*†] **Tim Bonnici** [1,2†] **Thomas Keen** [1] **Watjana Lilaonitkul** [1] **Mark J White** [4] **Nel Swanepoel** [3]

[1] *Institute of Health Informatics, University College London, London, UK*
[2] *Department of Critical Care, University College London Hospital, London, UK*
[3] *Centre for Advanced Research Computing, University College London, London, UK*
[4] *Digital Healthcare, University College London Hospital, London, UK*

Correspondence*:
Corresponding Author
steve.harris@ucl.ac.uk

[†] These authors have contributed equally to this work and share first authorship

ABSTRACT

Machine Learning for Health (ML4H) has demonstrated efficacy in computer imaging and similarly self-contained digital workflows, but has failed to substantially impact routine clinical care. Digital maturity is no longer a barrier where Electronic Health Record Systems (EHRS) are widely adopted. ML4H falls short because it needs an infrastructure for development, deployment and evaluation within the healthcare institution. In this paper, we propose a design pattern called a Clinical Deployment Environment (CDE). We sketch the five pillars of the CDE: (1) real world development supported by live data where ML4H teams can iteratively build and test at the bedside (2) an ML-Ops platform that brings the rigour and standards of continuous deployment to ML4H (3) design and supervision by those with expertise in AI safety (4) the methods of implementation science that enable the algorithmic insights to influence the behaviour of clinicians and patients and (5) continuous evaluation that uses randomisation to avoid bias but in an agile manner. Our CDE is derived from our experience of delivering machine learning models to the bedside in a large academic teaching hospital, and aims to answer for ML4H the same challenge that translational medicine brought to bear for drug discovery

Keywords: translational medicine, machine learning, health informatics, ml-ops, safety, artificial intelligence

INTRODUCTION

Bold claims and huge investments suggest Machine Learning (ML) will transform healthcare.(1) High impact publications showcase precision models that predict sepsis, shock, and acute kidney injury.(2, 3, 4) Outside healthcare, tech titans such as AirBnB, Facebook, and Uber create value from ML despite owning ’no property, no content and no cars’.(5) Inspired by this, and very much aware of the flaws and unwarranted variation in human decision making(6), government and industry are now laying heavy bets on ML for Health (ML4H).(7, 8)

Widespread adoption of electronic health records (EHR) is a prerequisite for this ambition. Yet while EHR adoption is growing at pace(9), those ML4H models that have reached the market rarely use the EHR. They are instead embedded in isolated digital workflows (typically radiology) or medical devices.(10) Here the deployment environment is either fully specified (devices), or static and self-contained (imaging). The problems are conveniently constrained by biology or a physical reality.

In contrast, the EHR is in constant flux. Both the data and the data model are updating. New wards open, staffing patterns are adjusted and from time to time major incidents (even global pandemics) disrupt everything. There are multiple interacting users, and eventually there will be multiple interacting algorithms, and we will have to tackle the ML equivalent of poly-pharmacy.(11) Algorithms will need ongoing care and attention. Whilst the aformentioned prediction models are developed on real-world data, this is insufficient. Despite their innovation and foresight, they find themselves in Gartner's trough of disillusinionment and at the bottom of the AI chasm.(12, 13) They have in reality not left the laboratory bench.

A future that sees ML4H generate value from the EHR requires more than a Trusted Research Environment (TRE) holding real-world data. TREs excel at the meeting the needs of population health scientists but they do not have the full complement of features required to take an ML4H algorithm from bench-to-bedside. Using drug development as an an analogy, a TRE is custom made for drug discovery not translational medicine.(14)

In this paper, we describe the functional requirements for a Clinical Deployment Environment (CDE) for ML4H. These differ from the classical components of translational medicine in that algorithms will require ongoing stewardship even after a successful deployment. The CDE is an infrastructure that manages algorithms with the same regard that is given to medicines (pharmacy) and machines (medical physics). Moreover, the value of ML4H will not just be from externally developed blockbuster models, but will also derive from specific and local solutions. Our vision of a CDE therefore enables both *development* and *deployment*.

Our CDE is supported by five pillars:

1.Real World Development
2.ML-Ops for Health
3.Responsible AI in practice
4.Implementation science
5.Continuous evaluation

We describe these pillars below alongside figures and vignettes reporting early local experience in building a CDE.

1 REAL WORLD DEVELOPMENT

Real-world data (RW-Data) means the use of observational data at scale augmented by linking across multiple data sources to generate insights simply not available from isolated controlled clinical trials.(15) The FDA uses data from tens of millions of patients in its Sentinel programme[1] to monitor drug safety, and the OpenSafely[2] programme in the UK generated impactful insights into COVID-19 within the first few months of the global pandemic.(16)

[1] <https://www.sentinelinitiative.org>

[2] <https://www.opensafely.org>

Given the sensitive nature of health data, these initiatives depend on expanding investment into Trusted Research Environments (TRE). (17) Data flows from source (primary, secondary, social care and elsewhere) to a single secure landing zone where research teams write the code to link, clean and analyse the data. The insights return to the bedside through clinical guidelines and policy. This offline '*data-to-code*' approach is also the dominant design pattern in ML4H projects but is fundamentally flawed.

A data-to-code feedback loop is measured in weeks and months, but ML4H interventions are virtual not physical, and must act by altering the behaviour of clinicians by providing insights for better decisions. As such, perfect information provided to the wrong person, or at the wrong moment, cannot be impactful. Excellent offline model performance provides no guarantee of bedside efficacy. Algorithms with inferior technical performance may even provide greater bedside utility.(18, 19)

ML4H requires instead a 'code-to-data' paradigm, and the first pillar of the CDE is the equivalent of an *internal* TRE *within* the healthcare institution rather than separate and external.(20) This permits live Real World Development (RW-Dev) that includes the end-user in rapid-cycle build-test-learn loops.

RW-Dev has four functional sub-requirements that distinguish it from a TRE. (1) Firstly, data updates must match the cadence of clinical decision making. For most inpatient and acute care pathways, decisions are in real-time (minutes or hours) at the bedside or in the clinic. (2) Secondly, experimental work on live data must be sandboxed and unable to harm the live clinical system (3) Thirdly, privacy must be managed such that teams are able to develop end-user applications that inevitably display patient identifiable information (PII) alongside the model outputs: an anonymous prediction is of little use to a clinician. (4) Fourthly, attention must be paid to developer ergonomics. Whether development and deployment are separated physically (the TRE paradigm) or functionally (languages and technologies), responsibilities divide between different teams. One team prepares the raw data and develops the model, and another prepares the live data and deploys the model. Ideally, the same team should be able develop and deploy both by bringing the CDE within the hospital, and secondly by aligning technology. This should accelerate iteratation, reduce cost and increase quality.(21)

We illustrate this idea with a description of our local real-world development platform in Figure 1, and provide an extended description in the Electronic Supplementary Material.

2 ML-OPS (FOR HEALTH)

Hitherto in ML4H, the data and the algorithm have been the 'celebrity couple'. State-of-the-art models trained on RW-Data deliver high impact publications.(4, 3) But only a tiny handful (just 8 studies in a recent high quality systematic review of 1909 ML4H publications(22)), are prospectively implemented. The standard offline data-to-code paradigm described above incurs a significant but 'hidden technical debt' that includes configuration, data collection and verification, feature extraction, analysis and process tools, compute and storage resource management, serving infrastructure, and monitoring.(23) In fact, the code for the underlying ML model is estimated to be at most 5% of the total code with the other 95% representing 'glue code' used to make the system work with generic packages. 'Glue-code', 'pipeline jungles', and 'dead experimental codepaths' are some of the anti-patterns that make the transition into production costly and hazardous.[3]

[3] One infamous example from the financial services sector saw a firm lose $170,000 per second (more than $400m in 45 minutes) when an outdated piece of code leaked into production. The firm in question was fined a further $12m for "inadequate safeguards" allowing "millions of erroneous orders".(24)

Agencies such as the FDA[4], EMA[5], and MHRA[6] are working toward safety standards for AI and machine learning, but the majority of these efforts derive from medical devices regulation. Treating Software as a Medical Device (SaMD) is appropriate where the algorithms operate within a constant and predictable environment (e.g. code embedded within a cardiac pacemaker). But, as already argued, ML4H models working with the EHR are likely to find themselves operating in a significantly more complex landscape. This inconstant environment where algorithms themselves may only have temporary utility has parallels to the commercial environment exploited so successfully by the tech giants.

These companies have cultivated an approach to model deployment called 'ML-Ops'. This combines the practices of 'DevOps' (a portmanteau of Software Development plus IT operations)(21) that focuses on the quality and speed with which software updates move from concept to production, with robust data engineering and machine learning. A typical ML-Ops system monitors raw input data, checks for distribution drift, provides a feature store to avoid train/serve skew and facilitate collaboration between teams, and maintains an auditable and monitored model repository.(25)

This constant adjustment of algorithm based on their continuously measured quality and performance needs a workforce as well as a technology stack. Just as the safe delivery of medicines to the bedside is the central activity of a hospital pharmacy team, the safe delivery of algorithms will require the development of similarly skilled and specialised practitioners, and we should expect to see clinical ML-Ops departments in the hospital of the future.

3 RESPONSIBLE AI IN PRACTICE

Pillars 1 and 2 should engender well designed and well engineered algorithms, but they do not protect against the unintentional harm that AI may induce. Algorithms can only learn from a digital representation of the world that in turn cannot encode moral or ethical standards. Unfair outcomes, discrimination against sub populations and bias are all reported shortcomings.(26) In a dynamic setting, risk can also arise in the form of degraded predictive performance over time. Models that modify clinician's behaviour alter patient profiles by design, but predictive success today inevitably erodes future performance by rendering obsolete the historical patterns that drove the performance of the original model.(27) Responsible AI in practice requires a systems approach that pre-empts and safe-guards against these potential risks to patients. We highlight three promising responses to components of this challenge that need to become part of the risk management approach for ML4H.

3.1 Model explainability

At the model selection stage, model explainability needs to be prioritised as one of the key metrics. Most AI models are not designed with explainability constraints and operate as 'black-box models'. On a practical level, 'black-box models' are unsuitable for healthcare because they pose risk scenarios where problems that occur can remain masked and therefore undetectable and unfixable. Explainable AI (XAI) research (28, 29, 30, 31) provides methods to highlight decision-relevant parts of AI representations and to measure and benchmarking interpretability. (32, 33) This is necessary for designing risk management as it enables a systematic interrogation of the trade-off between interpretability, model accuracy and the risk of model misbehaviour.

[4] <https://www.fda.gov>

[5] <https://www.ema.europa.eu/en>

[6] <https://www.gov.uk/government/organisations/medicines-and-healthcare-products-regulatory-agency>

3.2 Model fail-safes

Fail safes should be designed into support systems to pre-empt and mitigate model misbehaviour. The European Commission High-Level Expert Group on AI presented the Ethics Guidelines for Trustworthy Artificial Intelligence in April 2019 with recommendations for AI-support systems that continue to maintain human-agency via a human-in-the-loop oversight. Prediction models that map patient data to medically meaningful classes are forced to predict within the predetermined set of classes without the option to flag users when the model is unsure of an answer. To address this problem, there is good evidence that methods such as Bayesian deep learning and various uncertainty estimates (34) can provide promising ways to detect and refer data samples with high probability of misprediction for human expert review (35, 36, 37). This may even permit less interpretable models to operate when implemented in conjunction with an effective fail-safe system.

3.3 Dynamic model calibration

As discussed, models that influence the evolution of its own future input data are at risk of performance deterioration over time due to input data shifts. On the other hand, models can propose poor decisions as a result of the inherent biases found within the original dataset. In both cases, continual learning via model recalibration is required but continual learning remains a challenging paradigm in AI. Recalibration with non-stationary incremental data can lead to catastrophic forgetting when the new data negatively interferes with what the model has already learned(38), or a convergence where the model just predicts its own effect.(27) Here Pillar 1 (RW-dev) with suitable audit and monitoring via Pillar 2 (ML-Ops) will be required to overcome what would otherwise be a learning process encumbered by regulatory barriers.(39)

4 IMPLEMENTATION SCIENCE

Unlike medications, algorithms have no external existence. They can only impact health by influencing the behaviour of clinicians and patients. This corresponds to the second (T2) arm of translational medicine: implementation science.(14) A well designed, safe, and responsible AI algorithm may still be ineffective if it does not reach a modifiable target on the clinical pathway.(18) Implementation requires a multi-disciplinary approach including human-computer interaction, behavioural science, and qualitative analysis.(40)

We strongly argue that this task will be more difficult if done offline and in isolation. Pillar 1 crucially permits not just technical performance tuning of the algorithm but rapid build-test-learn cycles that directly involve the target user and the clinical pathway in question. This approach will reduce costs and improve impact, and inevitably lead to trade-offs that will appear surprising to those developing away from the bedside.(19, 11) This efficiency will again depend on the problem space: where the algorithmic target depends on information arising from the EHR rather than an isolated device or image, and where the pathway involves multiple end-users, then successful implementation will be near impossible if done sequentially (development then deployment) rather than iteratively.(40, 41) Academic health science centres must become design ’laboratories’ where rapid prototyping (build-test-learn) at the bedside crafts the deployment pathway for effectiveness rather than just efficacy.

Investigations to define how then system can influence behaviour will need specialist support and tooling. This might require tools embedded within the user interface to evaluate and monitor user interaction, and capture user feedback(42), or directed implementation studies.(43)

Despite the oft cited risks of alert fatigue with Clinical Decision Support Systems (CDSS)(44), there is good evidence that well designed alerts can be impactful.(45, 46, 47) Overt behavioural modifications will need a mechanism to explain their recommendation (as per XAI) or generate trust (see Pillar 5).(48) Trust will possibly be more important where behaviour modification is indirect through non-interruptive techniques (e.g. re-ordering preference lists or otherwise adapting the user interface to make the recommended choice more accessible).

5 CONTINUOUS CLINICAL EVALUATION

Our analogy with translational medicine breaks down at the evaluation stage. For drug discovery, evaluation is via a randomised controlled trial (RCT). Randomisation handles unanticipated bias and ML4H should hold itself to the same standard but of 350,000 studies registered on ClinicalTrials.gov[7] in 2020, just 358 evaluated ML4H, and only 66 were randomised.(49) As usual for ML4H, those RCTs were not interacting with EHR data. They were evaluations of algorithms supporting imaging, cataract screening, colonoscopy, cardiotocographs and more.(50, 51, 52, 53, 54, 55, 56)

Where the ML4H intervention delivers a novel biological treatment strategy, then it is appropriate to reach for the full paraphenalia used in Clinical Trials of Investigational Medicinal Products (CTIMPs).(2) But in many cases, algorithms will be used to optimise operational workflows and clinical pathways. These pathways may be specific and contextual rather than generalisable. Poor external validity is not a critique: the algorithm that is useful or important in one institution does not have to relevant in the next (the 'myth of generalisability').(57) Moreover, the algorithm is not the same as the patented and fixed active ingredient in a medicinal product. This is no single point in time nor single host environment at which it can be declared enduringly effective. This means that institutions deploying and relying on these tools need a strategy for rapid continuous clinical and operational evaluation.

This time the EHR may provide an advantage instead of just additional complexity. Since ML4H algorithms must be implemented through some form of direct or indirect CDSS, then the next logical step is to randomise the deployment of those alerts. This in itself is not novel. Randomised deterministic alerts from CDSS are part of the standard evaluation toolkit for quality improvement initiatives in at NYU Langone(58), and for research elsewhere.(59) At NYU Langone, such tooling permitted a small team to deliver 10 randomised trials within a single year.(58)

The final pillar in our CDE uses the same approach for the probabilistic insights derived from ML4H. Excellent and senior patient and public involvement, and ethical guidance, will be required to distinguish those algorithms that require per patient point-of-care consent from those that can use opt-out or cluster methods. But we think that latter group is large for two reasons. Firstly, patients are exposed to varying treatment regimes by dint of their random interaction with different clinicians based on geography (the healthcare provider they access) and time (staff holidays and shift patterns etc.). This routine variation in practice is summarised as the 60-30-10 problem: 60% of care follows best practice; 30% is wasteful or ineffective and 10% is harmful.(6) Secondly, because the intervention is informational, there is ethical precedent for patient level randomisation without consent (e.g. Acute Kidney Injury alerts).(59)

At our own institution, we have extended this ethical and safety case one step further, and we are piloting a study design where the randomisation is non-mandatory: a nudge not an order.(60) The clinician is explicitly invited to only comply with the randomisation where they have equipoise themselves. Where they have a preference, they overrule the alert (see Vignette 1 in the Electronic Supplemental Material).

[7] <ClinicalTrials.gov>

Embedded randomised digital evaluation should permit rapid evidence generation, and build the trust needed to support the implementation described under Pillar 4.

DRUG DISCOVERY PARALLELS

We have described a template for a Clinical Deployment Environment that supports the translation of ML4H algorithms from bench to bedside. Although the requirements differ, the objective is similar to that for drug development.

Most ML4H that derives value from the EHR is in the pre-clinical phase. In drug development, the objective of this phase is to identify candidate molecules which might make effective drugs. Evaluation is conducted in vitro. Metrics used to evaluate candidates, such binding affinity or other pharmacokinetic properties, describe the properties of the molecule.(61) For ML, the objective is to identify candidate algorithms, comprising of input variables and model structures, which might make the core of an effective CDSS. Evaluation is conducted offline on de-identified datasets. Metrics used to evaluate candidates, such Area Under the Receiver Operator Curve (AUROC), the F1 score and calibration, describe the properties of the algorithm.(62)

Phase 1 drug trials are the first time a drug candidate is tested in humans. They are conducted in small numbers of healthy volunteers. The aim of the trial is to determine the feasibility of progressing to trials in patients by determining drug safety and appropriate dosage. Drug formulation, the processes by which substances are combined with the active pharmaceutical ingredient to optimise the acceptability and effective delivery of the drug, is also considered at this stage. Phase 1 ML4H trials are the first time an algorithm candidate is tested within the healthcare environment. The aim of the trial is to determine the feasibility of progressing to trials of efficacy by ensuring the algorithm implementation is safe, reliable and able to cope with real-world data quality issues. The development of a mechanism to deliver of algorithm outputs embedded in the clinical workflow is also be considered at this stage.

Phase 2 drug trials involve recruitment of small numbers patients with the disease of interest, typically 50 – 200. The aim is to determine drug efficacy at treating the disease. Treating clinicians are involved in so far as they must agree to prescribe the drug for their patients. The trials are often too short to determine long term outcomes, therefore surrogate measures such biomarker status or change in tumour size are used as endpoints.(63) Phase 2 ML4H trials involve recruitment of small numbers of clinicians making the decision of interest, typically 5 – 10. The aim is to determine the efficacy of the algorithm in improving their decisions. Patients are involved in so far as they must agree to be on the receiving end of these supported decisions and identifiable data is required. Endpoints are markers of successful task completion in all cases. Investigations to determine ways in which the system could be more successful in influencing user behaviour are carried out at this stage. These include usability analyses, considerations of how well the ML4H/CDSS is integrated into the overall system and implementation studies to identify how best to optimise end-user adoption and engagement.(43)

Phase 3 drug trials involve the recruitment of large numbers of patients to determine whether a drug is effective in improving patient outcomes. The gold standard of trial design is a double-blinded randomised controlled trial (RCT). Phase 3 ML4H trials will require integration of data from multiple centres for algorithms acting on specific decisions but inevitably adapted to their local data environment.

CONCLUSION

Even this analogy stops short of the full task of deployment. With drug development, the universities and the pharmaceutical industry go on to take advantage of a supply chain to deliver the drug to the hospital with the necessary quality control and monitoring. Those prescribing and administering the drug have spent years in training, and are supported by pharmacists and medication safety experts. And even after the drug is administered, observation and long term follow-up continue to identify side-effects and long term hazards.

That network of expertise and infrastructure is largely in place where software *is* (not as) a medical device, but is only just being envisioned where the data driving ML4H comes from the EHR. This distinction needs to be made else the disillusionment with the promise of ML4H will continue. The technology does have the potential to change how we deliver health but the methodology alone is insufficient. The impressive demonstrations of the power of AI and ML to beat humans in games, and predict protein structures does not mean that these tools are ready for wide spread deployment.

But we should not be pessimistic. As per author William Gibson, it is clear that ”The Future Has Arrived — It’s Just Not Evenly Distributed.” Beyond healthcare, machine learning has already demonstrated that it can create reliably create value.(5) It is now our responsibility to take those lessons and adapt them for our patients.

The Five Pillars outlined here are a sketch of that redistribution. They are born from our local experience (Pillars 1, 2 and 5) and our wider observations (Pillars 3 and 4). They fundamentally are an argument for a professionalisation of ML4H. We envision a future where each algorithm is managed in a digital pharmacy with the same rigour that we apply to medicines. But unlike drugs, some of these algorithms will have their entire lifecycle, from development to deployment, managed by the local healthcare provider. Computer vision tasks that support diagnostic radiology can be partially developed offline. Components of sepsis prediction tools will transfer from institution to institution but will need adapting to local clinical workflows. But there will be opportunity and value for ML4H to optimise operational tasks that are temporary or specific to that institution. This means that some development and much of the deployment will require a suitably trained workforce, and an infrastructure perhaps supported by these five pillars.

CONFLICT OF INTEREST STATEMENT

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication.

FUNDING

SH is supported by a Health Foundation Improvement Science Fellowship, and by funding from the Biomedical Research Centre based at the University College Hospitals (UCLH) National Health Service (NHS) Foundation Trust and University College London (UCL). NS is supported by funding from Health Data Research UK. SH, TK, and NS are supported funds from the National Institute for Health Research

(Artificial Intelligence, Digitally adapted, hyper-local real-time bed forecasting to manage flow for NHS wards, AI AWARD01786) and NHS-X. WL is supported by a UKRI Ernest Rutherford Fellowship. The views expressed in this publication are those of the author(s) and not necessarily those of the National Institute for Health Research, NHS-X or the Department of Health and Social Care.

REFERENCES

1 .Bunz M, Braghieri M. The AI doctor will see you now: Assessing the framing of AI in news coverage. *AI & SOCIETY* **37** (2022) 9–22. doi:10.1007/s00146-021-01145-9.

2 .Komorowski M, Celi LA, Badawi O, Gordon AC, Faisal AA. The Artificial Intelligence Clinician learns optimal treatment strategies for sepsis in intensive care. *Nature Medicine* **24** (2018) 1716–1720. doi:10.1038/s41591-018-0213-5.

3 .Hyland SL, Faltys M, Hüser M, Lyu X, Gumbsch T, Esteban C, et al. Early prediction of circulatory failure in the intensive care unit using machine learning. *Nature Medicine* **26** (2020) 364–373.

4 .Tomašev N, Glorot X, Rae J, Zielinski M, Askham H, Saraiva A, et al. A clinically applicable approach to continuous prediction of future acute kidney injury. *Nature* **572** (2019) 116–119.

5 .McRae H. Facebook, Airbnb, Uber, and the unstoppable rise of the content non-generators. *The Independent* (2015).

6 .Braithwaite J, Glasziou P, Westbrook J. The three numbers you need to know about healthcare: The 60-30-10 challenge. *BMC medicine* **18** (2020) 1–8.

7 .The national strategy for AI in health and social care (2022).

8 .Digital future index 2021-2022 (2021).

9 .Everson J, Rubin JC, Friedman CP. Reconsidering hospital EHR adoption at the dawn of HITECH: Implications of the reported 9% adoption of a “basic” EHR. *Journal of the American Medical Informatics Association* **27** (2020) 1198–1205. doi:10.1093/jamia/ocaa090.

10 .Muehlematter U, Daniore P, Vokinger K. Approval of artificial intelligence and machine learning-based medical devices in the USA and Europe (2015-20): A comparative analysis. *The Lancet. Digital health* **3** (2021) e195–e203.

11 .Morse K, Bagley S, Shah N. Estimate the hidden deployment cost of predictive models to improve patient care. *Nature medicine* **26** (2020) 18–19.

12 .Steinert M, Leifer L. Scrutinizing gartner’s hype cycle approach. *Scrutinizing Gartner’s Hype Cycle Approach* (2010), 1–13.

13 .Keane PA, Topol EJ. With an eye to AI and autonomous diagnosis. *npj Digital Medicine* **1** (2018) 40, s41746–018–0048–y. doi:10.1038/s41746-018-0048-y.

14 .Woolf SH. The Meaning of Translational Research and Why It Matters. *JAMA* **299** (2008). doi:10.1001/jama.2007.26.

15 .Corrigan-Curay J, Sacks L, Woodcock J. Real-World Evidence and Real-World Data for Evaluating Drug Safety and Effectiveness. *JAMA* **320** (2018) 867. doi:10.1001/jama.2018.10136.

16 .Williamson E, Walker A, Bhaskaran K, Bacon S, Bates C, Morton C, et al. Factors associated with COVID-19-related death using OpenSAFELY. *Nature* **584** (2020) 430–436.

17 .Data research infrastructure landscape: A review of the UK data research infrastructure (2021).

18 .The DECIDE-AI Steering Group. DECIDE-AI: New reporting guidelines to bridge the development-to-implementation gap in clinical artificial intelligence. *Nature Medicine* **27** (2021) 186–187. doi:10.1038/s41591-021-01229-5.

19 .Shah NH, Milstein A, Bagley SC PhD. Making Machine Learning Models Clinically Useful. *JAMA* **322** (2019) 1351. doi:10.1001/jama.2019.10306.

20 .Guinney J, Saez-Rodriguez J. Alternative models for sharing confidential biomedical data. *Nature biotechnology* **36** (2018) 391–392.

21 .DevOps (2022).

22 .Ben-Israel D, Jacobs WB, Casha S, Lang S, Ryu WHA, de Lotbiniere-Bassett M, et al. The impact of machine learning on patient care: A systematic review. *Artificial Intelligence in Medicine* **103** (2020) 101785.

23 .Sculley D, Holt G, Golovin D, Davydov E, Phillips T, Ebner D, et al. Hidden technical debt in machine learning systems. *Advances in neural information processing systems* **28** (2015) 2503–2511.

24 .SEC Charges Knight Capital With Violations of Market Access Rule. https://www.sec.gov/news/press-release/2013-222 (2013).

25 .John MM, Olsson HH, Bosch J. Towards MLOps: A framework and maturity model. *Towards MLOps: A Framework and Maturity Model* (IEEE), vol. 2021 47th Euromicro Conference on Software Engineering and Advanced Applications (SEAA) (2021), 1–8.

26 .Amodei D, Olah C, Steinhardt J, Christiano P, Schulman J, Mané D. Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565* (2016). doi:10.48550/arXiv.1606.06565.

27 .Liley J, Emerson S, Mateen B, Vallejos C, Aslett L, Vollmer S. Model updating after interventions paradoxically introduces bias. *International Conference on Artificial Intelligence and Statistics* (2021), 3916–3924.

28 .Gunning D, Stefik M, Choi J, Stumpf S, Yang G. XAI—Explainable artificial intelligence. *Science Robotics* **4** (2019).

29 .Mueller S, Hoffman R, Clancey W, Emrey A, Klein G. Explanation in human-AI systems: A literature meta-review, synopsis of key ideas and publications, and bibliography for explainable AI. *arXiv preprint arXiv:1902.01876* (2019).

30 .Vilone G, Longo L. Explainable artificial intelligence: A systematic review. *arXiv preprint arXiv:2006.00093* (2020).

31 .Linardatos P, Papastefanopoulos V, Kotsiantis S. Explainable ai: A review of machine learning interpretability methods. *Entropy. An International and Interdisciplinary Journal of Entropy and Information Studies* **23** (2020).

32 .Doshi-Velez F, Kim B. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608* (2017).

33 .Hoffman R, Mueller S, Klein G, Litman J. Metrics for explainable AI: Challenges and prospects. *arXiv preprint arXiv:1812.04608* (2018).

34 .Abdar M, Pourpanah F, Hussain S, Rezazadegan D, Liu L, Ghavamzadeh M, et al. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion* **76** (2021) 243–297.

35 .Leibig C, Allken V, Ayhan M, Berens P, Wahl S. Leveraging uncertainty information from deep neural networks for disease detection. *Scientific reports* **7** (2017) 1–14.

36 .Filos A, Farquhar S, Gomez A, Rudner T, Kenton Z, Smith L, et al. A systematic comparison of bayesian deep learning robustness in diabetic retinopathy tasks. *arXiv preprint arXiv:1912.10481* (2019).

37 .Ghoshal B, Tucker T. Estimating uncertainty and interpretability in deep learning for coronavirus (COVID-19) detection. *arXiv preprint arXiv:2003.10769* (2020).

38 .Parisi G, Kemker R, Part J, Kanan C, Wermter S. Continual lifelong learning with neural networks: A review. *Neural Networks* **113** (2019) 54–71.

39 .Lee CS, Lee AY. Clinical applications of continual learning machine learning. *The Lancet Digital Health* **2** (2020) e279–e281. doi:10.1016/S2589-7500(20)30102-3.

40 .Sendak M, Ratliff W, Sarro D, Alderton E, Futoma J, Gao M, et al. Real-world integration of a sepsis deep learning technology into routine clinical care: Implementation study. *JMIR medical informatics* **8** (2020) e15182.

41 .Connell A, Black G, Montgomery H, Martin P, Nightingale C, King D, et al. Implementation of a Digitally Enabled Care Pathway (Part 2): Qualitative Analysis of Experiences of Health Care Professionals. *Journal of Medical Internet Research* **21** (2019) e13143. doi:10.2196/13143.

42 .Yusop NSM, Grundy J, Vasa R. Reporting Usability Defects: A Systematic Literature Review. *IEEE Transactions on Software Engineering* **43** (2017) 848–867. doi:10.1109/TSE.2016.2638427.

43 .Sutton RT, Pincock D, Baumgart DC, Sadowski DC, Fedorak RN, Kroeker KI. An overview of clinical decision support systems: Benefits, risks, and strategies for success. *npj Digital Medicine* **3** (2020) 1–10. doi:10.1038/s41746-020-0221-y.

44 .Phansalkar S, van der Sijs H, Tucker A, Desai A, Bell D, Teich J, et al. Drug-drug interactions that should be non-interruptive in order to reduce alert fatigue in electronic health records. *Journal of the American Medical Informatics Association : JAMIA* **20** (2013) 489–493.

45 .Park G, Kang B, Kim S, Lee J. Retrospective review of missed cancer detection and its mammography findings with artificial-intelligence-based, computer-aided diagnosis. *Diagnostics (Basel, Switzerland)* **12** (2022) 387.

46 .Sayres R, Taly A, Rahimy E, Blumer K, Coz D, Hammel N, et al. Using a Deep Learning Algorithm and Integrated Gradients Explanation to Assist Grading for Diabetic Retinopathy. *Ophthalmology* **126** (2019) 552–564. doi:10.1016/j.ophtha.2018.11.016.

47 .Main C, Moxham T, Wyatt JC, Kay J, Anderson R, Stein K. Computerised decision support systems in order communication for diagnostic, screening or monitoring test ordering: Systematic reviews of the effects and cost-effectiveness of systems. *Health Technology Assessment* **14** (2010). doi:10.3310/hta14480.

48 .McCoy L, Brenna C, Chen S, Vold K, Das S. Believing in black boxes: Machine learning for healthcare does not need explainability to be evidence-based. *Journal of clinical epidemiology* **142** (2022) 252–257.

49 .Zippel C, Bohnet-Joschko S. Rise of clinical studies in the field of machine learning: A review of data registered in ClinicalTrials.gov. *International journal of environmental research and public health* **18** (2021) 5072.

50 .INFANT CG. Computerised interpretation of fetal heart rate during labour (INFANT): A randomised controlled trial. *Lancet (London, England)* **389** (2017) 1719–1729.

51 .Titano JJ, Badgeley M, Schefflein J, Pain M, Su A, Cai M, et al. Automated deep-neural-network surveillance of cranial images for acute neurologic events. *Nature Medicine* **24** (2018) 1337–1341. doi:10.1038/s41591-018-0147-y.

52 .Wang P, Berzin T, Glissen Brown J, Bharadwaj S, Becq A, Xiao X, et al. Real-time automatic detection system increases colonoscopic polyp and adenoma detection rates: A prospective randomised controlled study. *Gut* **68** (2019) 1813–1819.

53 .Wu L, Zhang J, Zhou W, An P, Shen L, Liu J, et al. Randomised controlled trial of WISENSE, a real-time quality improving system for monitoring blind spots during esophagogastroduodenoscopy. *Gut* **68** (2019) 2161–2169.

54 .Lin H, Li R, Liu Z, Chen J, Yang Y, Chen H, et al. Diagnostic efficacy and therapeutic decision-making capacity of an artificial intelligence platform for childhood cataracts in eye clinics: A multicentre randomized controlled trial. *EClinicalMedicine* **9** (2019) 52–59.

55 .Turakhia MP, Desai M, Hedlin H, Rajmane A, Talati N, Ferris T, et al. Rationale and design of a large-scale, app-based study to identify cardiac arrhythmias using a smartwatch: The Apple Heart Study. *American Heart Journal* **207** (2019) 66–75. doi:10.1016/j.ahj.2018.09.002.

56 .Long E, Lin H, Liu Z, Wu X, Wang L, Jiang J, et al. An artificial intelligence platform for the multihospital collaborative management of congenital cataracts. *Nature Biomedical Engineering* **1** (2017).

57 .Futoma J, Simons M, Panch T, Doshi-Velez F, Celi LA. The myth of generalisability in clinical research and machine learning in health care. *The Lancet Digital Health* **2** (2020) e489–e492. doi:10.1016/S2589-7500(20)30186-2.

58 .Horwitz LI, Kuznetsova M, Jones SA. Creating a Learning Health System through Rapid-Cycle, Randomized Testing. *New England Journal of Medicine* **381** (2019) 1175–1179. doi:10.1056/NEJMsb1900856.

59 .Wilson FP, Martin M, Yamamoto Y, Partridge C, Moreira E, Arora T, et al. Electronic health record alerts for acute kidney injury: Multicenter, randomized clinical trial. *BMJ* (2021) m4786. doi:10.1136/bmj.m4786.

60 .Wilson MG, Asselbergs FW, Harris SK. Learning from individualised variation for evidence generation within a learning health system. *British Journal of Anaesthesia* (2022).

61 .Lo YC, Rensi SE, Torng W, Altman RB. Machine learning in chemoinformatics and drug discovery. *Drug Discovery Today* **23** (2018) 1538–1546. doi:10.1016/j.drudis.2018.05.010.

62 .Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): The TRIPOD Statement. *Annals of Internal Medicine* **162** (2015) 55–63. doi:10.7326/M14-0697.

63 .Van Norman GA. Phase II Trials in Drug Development and Adaptive Trial Design. *JACC: Basic to Translational Science* **4** (2019) 428–437. doi:10.1016/j.jacbts.2019.02.005.

FIGURES

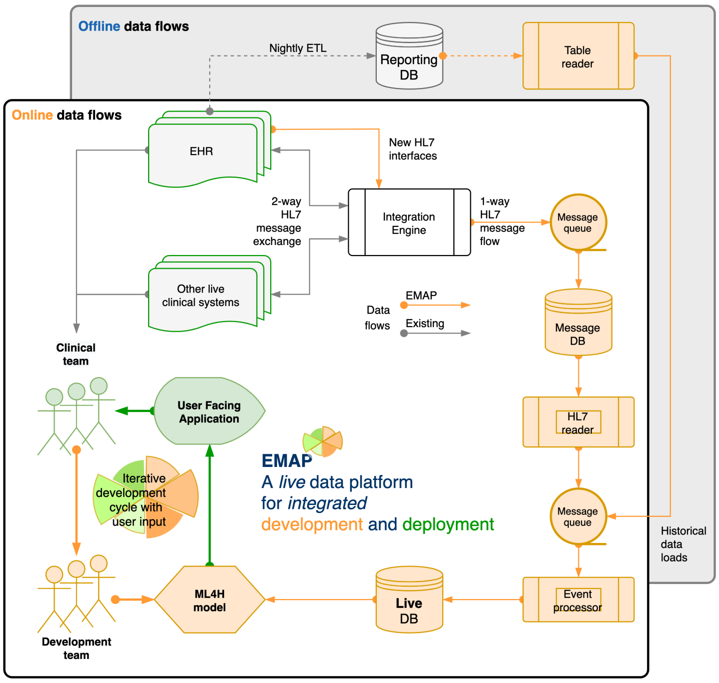

Figure 1. Our real-world development is performed on the Experimental Medicine Application Platform (EMAP). EMAP is a clinical laboratory within which ML4H researchers can iteratively build, test and gather feedback from the bedside. It unifies the data and the tools for off-line and online development of ML4H models. In brief, EMAP builds a patient orientated SQL database from HL7 version 2 (HL7v2) messages that are being exchanged between hospital systems. HL7v2 messages are ubiquitous in health care, and the *de facto* standard for internal communication.

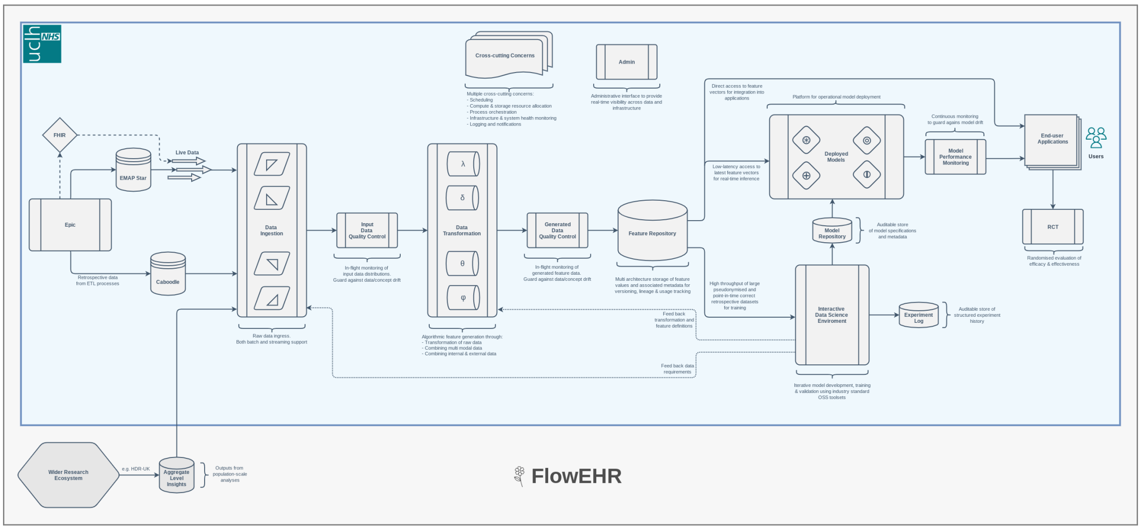
ucl h NHS
Cross-cutting Concerns
Multiple cross-cutting concerns:
- Scheduling
- Compute & storage resource allocation
- Process orchestration
- Infrastructure & system health monitoring
- Logging and notifications
Admin
Administrative interface to provide real-time visibility across data and infrastructure
FHIR
Live Data
EMAP Star
Epic
Retrospective data from ETL processes
Caboodle
Data Ingestion
Raw data ingress. Both batch and streaming support
Input Data Quality Control
In-flight monitoring of input data distributions. Guard against data/concept drift
Data Transformation
λ
δ
θ
φ
Algorithmic feature generation through:
- Transformation of raw data
- Combining multi modal data
- Combining internal & external data
Generated Data Quality Control
In-flight monitoring of generated feature data. Guard against data/concept drift
Feature Repository
Multi architecture storage of feature values and associated metadata for versioning, lineage & usage tracking
Direct access to feature vectors for integration into applications
Low-latency access to latest feature vectors for real-time inference
Platform for operational model deployment
Deployed Models
Continuous monitoring to guard agains model drift
Model Performance Monitoring
End-user Applications
Users
RCT
Randomised evaluation of efficacy & effectiveness
Model Repository
Auditable store of model specifications and metadata
High throughput of large pseudonymised and point-in-time correct retrospective datasets for training
Feed back transformation and feature definitions
Feed back data requirements
Interactive Data Science Enviroment
Iterative model development, training & validation using industry standard OSS toolsets
Experiment Log
Auditable store of structured experiment history
Wider Research Ecosystem
e.g. HDR-UK
Aggregate Level Insights
Outputs from population-scale analyses
FlowEHR

Figure 2. Our ML-Ops platform is called FlowEHR. Moving from left to right across the figure, the system monitors raw input data including checks for distribution shift, builds features with testable and quality controlled code, makes those features available to for both training and predictions to avoid train/serve skew, and maintains an auditable and monitored model repository.