Advancing Clinical Trials via Real-World Aligned ML Best Practices

Karen Sayal* **Finnian Firth** **Markus Trengove** **Lea Goetz**
Artificial Intelligence and Machine Learning, GSK
karen.k.sayal@gsk.com

Abstract

There is an increasing drive to integrate machine learning (ML) tools into the drug development pipeline, to improve success rates and efficiency in the clinical development pathway. The ML regulatory framework being developed is closely aligned with ML best practices. However, there remain significant and tangible practical gaps in translating best practice standards into a real-world clinical trial context.

To illustrate the practical challenges to regulating ML in this context, we present a theoretical oncology trial in which a ML tool is applied to support toxicity monitoring in patients. We explore the barriers in the highly regulated clinical trial environment to implementing data representativeness, model interpretability, and model usability.

1 Introduction

As regulators begin to set standards for ML deployment across a range of industries, there is an increasing need to address the operational gap between high-level guidance and the practical constraints of specific real-world use cases of ML. This paper illustrates the operational gaps which arise when applying potential regulatory guidance to a hypothetical real-world use case of ML in clinical trials. We focus on the EMA discussion paper on regulation in the AI-medicine's lifecycle. We outline a hypothetical case in which a ML tool is integrated into a clinical trial to predict patient risk of developing acute kidney injury (AKI).

Despite its potential for positive impact on patient care, the tool's levels of transparency, data representativeness, prospective validation, interpretability, data protection, and performance testing all fall short of the baseline standards set out by the EMA's proposal. Whilst ML can significantly improve the quality of patient care in clinical trials, our case vignette illustrates why it can be infeasible for ML to achieve the model and data requirements that regulators might expect. It is imperative to strike a context-sensitive balance that enables the beneficent value of adopting ML and maintains baseline standards for ML adoption. We propose this case as a thought experiment for the ML community to consider how to collaborate with regulators in an optimally balanced manner.

2 Clinical Trial Development

Clinical trials are designed to test whether a new type of treatment, the 'experimental therapy', is safe to administer to patients and if it is effective against treating a disease. The experimental therapy is often a newly designed medicine which has completed extensive testing in the laboratory. The clinical trial is often the first time in which the new chemical is given to humans. Therefore, clinical trials operate under strict regulatory processes to ensure the research is appropriate and that no undue harm is caused to trial participants.

Workshop on Regulatable Machine Learning at the 37th Conference on Neural Information Processing Systems (RegML @ NeurIPS 2023).

Clinical trials can be broadly classified into two main categories: early phase and late phase trials. Early phase trials, particularly Phase 1 trials, are primarily designed to evaluate a drug's safety profile. Early phase trials are usually conducted in a small group of patients and determine whether an experimental therapy can proceed to more extensive testing. Early phase trials in oncology are unique in that cancer patients, rather than healthy volunteers, are recruited at the phase 1 stage. Conversely, late phase trials involve larger patient cohorts and compare the experimental agent to treatments already established in routine clinical care.

Clinical trials are expensive, lengthy, and resource intensive. There is an increasing drive to effectively utilise ML tools to improve efficiency. A trial is more efficient when either the same clinical endpoint is reached in a shorter period, when fewer patients are recruited to offer the same insight into an experimental therapy, or when the safety monitoring process is streamlined. Such an outcome would be considered a successful improvement on clinical trial design.

3 Existing Regulatory Guidance

We propose our thought experiment in the context of the development of regulation for ML in clinical trials. Our thought experiment highlights a critical tension that regulators will have to navigate.

Legislation for AI/ML in general is currently nascent – the European Union's AI Act is currently one of the most advanced sector-agnostic proposed pieces of legislation and is still in negotiation at the time of writing [1]. In general, the purpose of legislation and regulation is to reduce the incremental risks that the adoption of ML introduces.

Since the risk profile of ML applications will differ between use cases, sectoral regulators are increasingly scoping out regulation for their sectors. The European Medicines Agency (EMA) has regulatory authority over medicine development in the EU. In July 2023, the EMA issued a reflection paper [2]. The reflection paper discusses the uses of ML across the medicine development lifecycle. Although it does not contain concrete suggestions for regulations, the paper outlines the standards of best practice in ML with a view towards codifying the standards into regulation for ML in drug development and clinical applications.

We agree with the aim and methodology of the discussion paper. It is imperative to set standards and procedures for regulators to verify the safety of ML adoption, particularly where it affects patients. It would be dangerous for practitioners to deploy ML without such standards or processes. It is also important to establish clear guidance for practitioners, so that they can adjust their expectations and practices accordingly.

However, the aim of our paper is to draw out a tension in this regulatory process. Whilst the discussion paper outlines agreeable standards of best practice for ML in drug development generally, these standards – when applied broadly across the medicine lifecycle — are so stringent that they risk prohibiting the applications of ML in domains of drug development in which adherence to these general standards would be impracticable or impossible.

Our paper proceeds by outlining one such domain – clinical trials — in which the adoption of ML could significantly improve patient welfare. The purpose of our paper is twofold: first, to suggest potential resolutions to this conflict through permitting justified and context-specific deviations from "gold standard" practices; second, to encourage the ML community to engage with this tension to find a productive balance between setting clear industry standards and processes, and remaining sufficiently flexible to permit welfare-improving ML applications.

4 Case Vignette

We present a theoretical study to highlight the practical gaps between existing regulations and typical ML workflows. We discuss the findings raised from a theoretical clinical trial in which twenty patients with advanced breast cancer receive an experimental therapy.

Clinical trials are designed and managed by a clinical development team, a specialist team of clinicians, nurses, pharmacists, statisticians, and trial managers who are commonly based in academic institutions or pharmaceutical companies. In our case vignette, the composition of the clinical development team is unusual in that it also includes ML engineers. During the trial, patients will receive the therapy from a specialist clinical trials unit in a UK-based hospital.

The aim of the trial is to assess the safety profile of the experimental therapy and identify a safe treatment dose for patients. The patients enrolled in the trial have exhausted all treatments available as part of standard clinical care. In particular, they have received multiple rounds of previous chemotherapy, which has reduced their baseline kidney function. Therefore, giving the experimental therapy as part of the clinical trial has the increased risk of causing further kidney damage. If the damage is significant, it may cause AKI. In extreme cases, AKI can be life-threatening.

The clinical development team decides to integrate an AKI monitoring tool into the planned trial. Recruited patients receive the experimental therapy as per standard trial practice. In addition to regular blood tests and clinical review, patients are reviewed with an AKI software monitoring tool. The anticipation is that the ML-based tool will enable better resource allocation, by alerting clinicians earlier to the onset of kidney damage which can be managed by simpler and less intense measures. As part of our case study, the selected tool will be an approved ‘Software as a Medical Device’ and will comply with the required software safety standards.

5 Data Representativeness

Data representativeness is one of the most important factors in any discussion about applying a model. In clear alignment with this principle, the EMA advises that “performance should be tested with prospectively generated data that is acquired in a setting or population representative of the intended context of use” [2]. However, due to the complexity of human data, it is often impossible to find perfectly representative datasets for a new clinical trial. In some cases, the gap between apparently similar datasets is too wide to bridge. But, in other cases, appropriate care can be taken to re-purpose a dataset. To enable this, we stress the need for a regulatory framework by which ML practitioners can prove the appropriateness of models trained on “imperfect” data.

Recall that in our case study, the clinical development team includes ML engineers; during trial design, these engineers explore possible models for predicting AKI risk. One of the highest-accuracy models found in the literature is a recurrent neural network (RNN) trained on a retrospective clinical dataset acquired from the US Department of Veteran Affairs (USDVA); for the reader’s reference, this example is loosely based on the tool and datasets published by Tomasev et al.[3]. In addition, the engineers develop their own random forest model, based on a similar existing model trained on a set of local hospital records.

The multilayer RNN initially seems promising, due to its extensive documentation, as well as the fact that it was successfully adopted for a previous clinical trial. However, the engineers note two major issues affecting the representativeness of the data collected.

Firstly, they note that the data is collected from US military veterans, a population strongly enriched for the male sex. In contrast, the trial is recruiting patients with breast cancer, a disease driven by uncontrolled oestrogen signalling, and therefore enriched in females. There is evidence that rates of AKI are higher in men compared to women [4]. If unaccounted for, this discrepancy could result in a significantly higher false positive rate for AKI risk on the trial. The engineers also note differences in the racial makeup of the US data against the theoretical UK patient base for the trial. This is more difficult to precisely quantify but could lead to other unintended shifts in model behaviour.

Secondly, they note that the clinical outcome in the data set is taken to be dialysis, an extreme interventive treatment to replace kidney function altogether, while the kidneys recover. Whilst appropriate for the USDVA’s use case, kidney injury requiring dialysis is already far too advanced for this clinical trial: a much earlier warning is necessary to ensure the experimental therapy is safe. Furthermore, some of the clinical thresholds referenced within the documentation do not match those commonly used in UK hospitals. Even if dialysis were an appropriate endpoint, the model would need to be re-trained against thresholds more appropriate for the trial at hand. The EMA proposal recommends using "*a priori* defined thresholds for performance metrics" [2], but as shown here, setting these metrics need to be context sensitive.

Concerned by their findings with this tool, the engineers also examine an in-house random forest model, which is currently being used to predict cardiac risk. Despite this dramatically different endpoint, the engineers point out several promising aspects of the model and the data it was trained on.

Firstly, the engineers note that although the model was trained to predict cardiac risk, the underlying data set contains a diverse and complete panel of clinical tests and bloodwork. In particular, all the necessary data to train against the appropriate AKI endpoint are present. With thorough documentation on how the original ML pipeline was prospectively validated, including normalization, handling of confounders, etc., it is immediately clear to the engineers how the pipeline could be easily adjusted to model renal risk, rather than cardiac.

Secondly, the engineers note that the data set was collected from patients at UK hospitals. The patient demography very closely aligns with the expected patient population for the trial. Furthermore, the size of the dataset would allow for an appropriately sampled female-skewed population, to mirror the likely design of the trial.

We turn to the question of what these ML engineers do now. Note that none of the data they found was a perfect match for the trial being designed: both deviated from the ideal problem statement, to varying degrees. The regulatory position on the first option should be clear: it is not sufficient. The second option is better but leaves unanswered questions. To which principles can they adhere to ensure work undertaken to train a new model would be sufficient from a regulatory perspective? Could actionable guidelines have helped or accelerated their initial task of evaluating these datasets and models? What evidence do regulators require to demonstrate the appropriateness of the tool?

We suggest that, although the examples here are simplistic for illustrative purposes, they reflect common issues with real-world data representativeness. Because these issues vary so significantly from problem to problem, we do not believe there can be a one-size-fits-all recommendation. Rather, we stress the need for general regulatory principles to inform and guide the practical assessment of data representativeness, but also for a framework in which individual trial teams can efficiently engage with regulators to resolve queries specific to their circumstances.

Finally, it is worth noting that non-regulatory-based progress is also possible. One possibility would be a coordinated effort to amass high-quality clinical datasets in a central repository, with caveats around data collected during the 2020 – 2022 pandemic. Clinical management pathways changed suddenly and significantly during the pandemic, creating bias specific to this period. Another possibility would be a suite of representation learning approaches to formally quantify the differences between different patient populations. Of course, these are both massive undertakings requiring prospective planning, and cannot be achieved by an ad-hoc collection of datasets and models from non-standardised sources. Additionally, the ultimate utility of these still relies on the relevant regulatory guidance and framework in place.

6 Model Interpretability

The EMA encourages the use of black box models only when ’interpretable models show unsatisfactory performance or robustness’ [2], and similar ideas are discussed in the ML community for various human-facing applications of ML. Interpretability is an important safety aspect of ML tools used in clinical trials. If, as in our case vignette, a model is generating a risk score, then the reviewing clinician will need to understand which input features are driving the prediction and make a clinical judgement on whether to accept or reject the model’s risk assessment. Furthermore, interpretability is a central feature for regulatory purposes. When an investigational agent is discontinued for a patient, clear documentation is required of the reasons for the decision to ensure it is a balanced decision within the overall trial framework.

In an ideal scenario, the most highly performant model is an interpretable model. However, in practice there is often a trade-off between model capacity and performance (e.g., deep neural networks) and ease of interpreting the model (e.g., linear regression). This presents a dilemma for the community: is it acceptable and appropriate for model interpretability to be available at the expense of model performance? If so, what level of diminution of performance is acceptable?

We believe that the ideal balance between model accuracy and model interpretability cannot be set by globally defined regulatory standards. Instead, this trade-off needs to be decided on a case-by-case basis by an interdisciplinary clinical development team, including at the very least clinicians and ML engineers. A real-world ready regulatory framework for ML in clinical trials would need to avoid generic requirements for interpretability, and instead allow justified deviations from idealistic "ML best practices", where appropriate. We encourage a conversation between the ML community, regulators, clinicians and AI ethicists on how the conflicting needs of high model performance and model interpretability should be balanced.

Notably, in cases where black box models must be used for performance considerations, the EMA advises integrating interpretability methods, such as SHAP and/or LIME analysis into black box models [2]. However, requiring such explicitly listed interpretability methods and metrics is overly prescriptive, and does not necessarily achieve the intended outcome of improving the safety and trustworthiness of ML tools in clinical trials. We caution against the indiscriminate use of interpretability methods to solely satisfy a regulatory requirement, rather than providing clinical insights. Furthermore, by allowing ML developers with domain expertise to choose the current state of the art interpretability methods, rather than prescribing specific methods/metrics that may soon be outdated, the EMA would future-proof their regulation.

6.1 Incorporating model interpretability into clinical trial design

One approach to improve the safety and accountability of ML tools in clinical trials that sidesteps model interpretability is the use of human-interpretable features on a model which may otherwise be considered "uninterpretable", such as a deep neural network. Human-interpretable features as model inputs, intermediate model representations, and/or model outputs will provide reassuring transparency for clinicians using the tool.

We propose that a well-designed clinical workflow around a model, such as using a human-in-the-loop to sanity check the model's use of human-interpretable features, may achieve the necessary interpretability and traceability of a clinical decision even with a black box model. It should also greatly increase the safety and trustworthiness of clinical ML tools, not just for clinicians but also for patients.

When appropriately supported by flexible regulatory guidance, an interdisciplinary clinical development team with both clinical and ML expertise will be best placed to select the optimal interpretability solution to ensure meaningful human oversight. Any interpretability tool should be identified prospectively and defined as part of the trial protocol in a manner aligned with individual trial requirements.

7 Additional Considerations for Real-World Usability

Real-world usability is too broad to be discussed in detail within a single article, but four specific points are worth mentioning that bear on the complexity of applying and regulating ML in the clinical trial context.

The first concerns the requirements for clinically stable performance of an ML model. In our vignette, clinically relevant physiological processes could have been represented in a multimodal fashion to avoid assay-dependent and clinician-dependent decision thresholds. It is important to avoid over-reliance on a small panel of blood tests, which may be vulnerable to spurious data artefacts and differences in clinical judgement. Instead, an expanded kidney-specific feature set could be constructed consisting of a wider panel of blood tests, urine output, measures of urine quality and kidney imaging scans. In this way, the ML model can better replicate the manner in which doctors conduct clinical assessments, since patients are reviewed and clinical management decisions made by gathering results from multiple complementary axes of information.

The second point concerns the logistical complexities of regulating a ML-integrated clinical trial, given the difficulties with introducing a novel tool to an established clinical framework. Integrating a large tool into a hospital compute system creates a logistical obstacle that exposes the performance of the tool to external risk. If, for instance, a hospital-wide update causes the model to fail, this would require a "bridge re-evaluation" per the EMA paper, but such a re-evaluation might be infeasible in the context of an active clinical trial and can result in the suspension of the clinical trial in the absence of a contingent clinical protocol. A possible solution here would be to define a non-ML based pathway as part of the trial protocol in the case of failure. Clinicians' management of the ML tool creates further logistical obstacles: if clinicians over-rely on the results of the tool, this impedes the holistic evaluation of the patient and the interpretability of the model results. To avoid this, clinicians with extensive trial experience can be assigned to interpret model results as a matter of trial protocol; ML and clinical development teams can collaborate to optimise model usability; the tool's user interface can be optimised for non-technical users; and generated reports can be structured to contain only information relevant to clinician review.

The third point concerns the risk-benefit profile of an ML tool for a clinical trial. In our vignette, the AKI tool for a clinical trial would be used by clinicians in a similar way to an AKI tool applied in the non-trial setting. The key distinction is that different clinical contexts have different risk-benefit profiles. The experimental therapy investigated as part of a clinical trial has an incomplete toxicity profile, whereas the toxicity profile of routinely used treatments is well-described. Therefore, patients in a clinical trial face greater uncertainty in their toxicity exposure, which could potentially be greater. The benefits offered by better toxicity monitoring when treatment side-effects are uncertain outweighs some of the potential risks associated with a new tool. The nuanced determination of risk-benefit profiles should be decided jointly by the ML engineers and clinicians using the tool. The clinicians can present short clinical case studies to their collaborating ML engineers, which would serve as a springboard from which they can jointly decide appropriate thresholds for sensitivity, specificity, positive predictive value and/or negative predictive value, relevant for the clinical context at hand. We believe this engagement between clinicians and engineers would be enforced through regulation and would be an iterative process, which should run throughout the ML development process.

The fourth point concerns clinical trial training requirements. The standards should be relevant and replicable across different countries and encompassed within the existing 'Good Clinical Practice (GCP) for clinical trials' training courses. All members of the clinical development team, including ML engineers, will be required to complete GCP training for regulatory compliance. There is also a statutory requirement to notify the supervising regulatory body of serious breaches of GCP or the trial protocol

8 Discussion

We have used a theoretical case study to illustrate some of the practical challenges in regulating ML in the context of clinical trials. Since the regulation in question is not yet in place, there are no real-world case studies that could serve as examples for a vignette. As a next best solution, we have based the case study on both existing ML-tools [3] and parameters that are common in oncology trials. The hypothetical nature of the case study is intended to focus attention on the general principles and challenges of ML deployment in clinical trials. Although the more detail-oriented reader will (somewhat understandably) be unsatisfied with the lack of methodological detail present within the examples, we strongly encourage them to continue to apply that rigor in the communal discussion we hope to spark.

Above all, our case vignette has demonstrated that the responsible deployment of ML-based tools in clinical trials must be informed by the unique context of the trial in question. This will allow us to find an appropriate trade-off between the conflicting values of strict adherence to ML best practices and a more pragmatic approach that aims to do the best for patients under suboptimal conditions. Trying to specify such context-sensitive trade-offs in general, via prescriptive regulation, would be prohibitively difficult. On the other hand, the primary purpose of a strict regulatory framework is to protect patients from ignorance, negligence, or even bad intentions; an overly flexible regulatory framework would fail to achieve this. This is particularly the case when decisions on ML practices lie in the hands of the tool's developers alone. Specifically: given the circumstances of a particular trial, ML engineers may occasionally need to adapt ML best practices. And, whilst it should remain possible for them to decide on an appropriate approach, it will be absolutely essential for an independent panel of ML experts, reporting to the regulatory authority, to ratify that approach. Furthermore, publishing the results and justifications of these reviews would provide transparency and could act as exemplars to guide clinical development teams and ML engineers in the development and deployment of future ML-based clinical tools.

Related legislative and regulatory developments already suggest one possible method for achieving the requisite balance between control and flexibility. For example, the National Institute for Standards in Technology (NIST) has published a Risk Management Framework (RMF) for the adoption of ML [5]. The RMF does not prescribe practical standards, but rather outlines a process for clarifying and mapping risks and mitigation strategies. The nascent US AI Accountability Bill, rather than prescribing general standards for ML, requires that providers develop reports on the risks and mitigation of their ML applications in consultation with experts and stakeholders, and that regulators assess these reports [6]. Similarly, the EU AI Act requires that providers of ML applications submit conformity assessments to regulatory bodies which detail the system's risk management system, identify and mitigate known and foreseeable risks, and record adequate testing and validation, considering these measures in light of the intended purpose and design choices in the development process [1].

These regulatory proposals suggest a set of possible mechanisms for balancing regulation and flexibility by standardizing a reporting and assessment process overseen by regulators. This process-oriented regulation allows regulators to make more granular judgments about the risk-benefit profile of a ML application. Such a mechanism can be deployed alongside a set of baseline standards providing a minimum threshold for responsible ML development. Moreover, as regulators review more applications, they can develop a body of precedent and principles that, like a body of case law, can give developers more certainty.

Below we discuss two examples to showcase how one may find appropriate trade-offs between regulatory requirements and flexibility in two very different contexts: late phase vs early phase clinical trials.

8.0.1 Example trade-off 1: interpretability in late phase clinical trials

Finding the right trade-off between model performance and model interpretability is a high priority in the late phase clinical trial setting, where it ensures the model's prediction is aligned with the established pathophysiological understanding of the disease process. Late phase trials, such as phase 3 trials, compare a standard routinely delivered clinical treatment with a novel treatment. If the novel treatment shows superior efficacy to the standard treatment during the trial, it will then become standard of care for all in-scope patients. Therefore, in such a high-stakes context, being able to interpret the model's prediction will be critical for developing clinician and regulatory confidence in the model's real-world utility.

What complicates the issue of interpretability in clinical applications of ML is that there are multiple stakeholder groups – researchers, ML engineers, legislators, clinicians and patients – all of which will likely need different types of explanation, appropriate for the nature of their expertise [7]. On the one hand, regulators may be primarily interested in gaining sufficient insight into a "black-box model" to perform a risk assessment of its application in a clinical trial. Clinicians, on the other hand, may want to identify which clinically relevant features drive model predictions.

Thus, when regulators ask for interpretability, it is unclear how different, potentially conflicting, interpretability needs should be met. A good compromise for legislators might be to allow the clinical development team to prioritize the interpretability needs of one stakeholder group over another, but to review the justification to do so on a case-by-case basis. For example, in late-stage clinical trials interpretability of predictive features may be essential for clinicians who need to take model outputs into account when making treatment decisions. Therefore, when developing the model, ML engineers need to focus on feature-based interpretability. Whether this is achieved through an inherently interpretable model, hand-crafted human-interpretable features or post-hoc feature-based methods such as LIME, should be a joint decision of developers and clinicians, balancing interpretability requirements with model performance where necessary.

In practice, this raises an interesting ML problem: different interpretability methods are often discordant with each other [8]. In contrast, different clinical features often represent the same underlying physiological system. For example, creatinine and eGFR are common blood tests used to monitor kidney function. From a clinician's perspective, they are equivalent features and are both equally valid ways to represent the urinary organ system. Specifically, an interpretability method which top ranks creatinine has the same meaning for a clinician as an interpretability method which top ranks eGFR. This equivalence of features for clinicians stands in stark contrast to the 'interpretability disagreement' [8] of different interpretability metrics. We encourage the ML community to engage with clinicians to learn about such issues and develop new methods that address them.

8.0.2 Example trade-off 2: representativeness in early phase trials

In the early phase setting, data representativeness is the primary challenge. Patient numbers in an early phase trial are low, generally in the range of 10 – 30 patients. Therefore, any ML model would ideally be trained and validated on a larger, non-trial dataset and then applied to the early phase trial context. It will be essential to assess the representativeness of the training dataset, including leveraging domain experts to ensure the training dataset can sufficiently capture the key physiological characteristics of the clinical trial population.

Because of the unique nature of clinical trial populations, the representativeness of training datasets will never be perfect and the use of domain adaptation (DA) methods to improve the representativeness between trial and non-trial clinical populations is yet to be demonstrated. The work on DA by Guo and colleagues [9] illustrates this challenge. Although their data is collected outside a clinical trial setting, the findings and principles are still highly relevant for our clinical trial case vignette.

Guo et al explored the well-known MIMIV-IV database from intensive care unit (ICU) patients at the same clinical unit in the same hospital, but at different historical periods (2008 – 2019). Patient cohorts were generated from consecutive three-year time blocks; clinical features such as blood test results, ICU patient monitoring charts and patient diagnoses were extracted; and four clinically meaningful prediction tasks, including the development of sepsis, were defined. To learn data-invariant properties for each task, domain generalisation (DG) and unsupervised domain adaptation (UDA) algorithms like including correlation alignment (CORAL), maximum mean discrepancy (MMD), domain adversarial learning (AL), invariant risk minimisation (IRM) and group distributionally robust optimisation (GroupDRO) were applied to a baseline deep neural network.

In all experiments, no DG or UDA method fully mitigated the temporal data drift and they did not improve model robustness [9]. Although the published results are from an ICU setting, the general principles and nature of application are similar to our clinical trial vignette. In both settings, standard clinical results with a focus on bloodwork are being used to predict the risk of future clinical deterioration of a patient. The factors accounting for the time- and location-dependent drift in clinical datasets are similar, with both types of drift arising due to evolutions in medical practice, changes in population demographics, and possibly differences in data collection and recording practices. Of these, changes in medical practice, such as healthcare professionals recognising sepsis and starting treatment earlier [10], are likely the largest contributor to clinical data drift.

The study underlines the complexity and nuances of clinical data and suggests that a straightforward ‘plug-and-play’ of DA methods would likely be ineffectual. Instead, regulators should outline general criteria by which DA methods can be safely and effectively used in clinical datasets. It is then up to the ML practitioner to decide on which specific DA method to apply and to demonstrate that the selected method fulfils the regulatory criteria.

The examples above demonstrate that regulations must be consistently applicable to the smallest groups and the largest companies. However, this consistency is difficult to achieve given the pace at which machine learning evolves. If guidelines are too specific, they will need near-constant revision to keep pace with the current ML landscape. If, on the other hand, they are too general, individual cases carry an undesirable ambiguity.

This is a tremendously difficult problem to solve, and no satisfactory solution currently exists, but we believe progress can only be made by attempting to find a middle ground. We need appropriately detailed principles by which data and models should be evaluated, but to allow flexibility-within-limits in the way these principles are proved. Correctly balanced principles can only be derived by a collaboration of clinicians, ML experts, and regulators. The same is true for defining the “rules of evidence": it should be clear upfront to all parties involved what needs to be proved, with any given approach. Although new model frameworks will emerge, requiring different specific considerations, we feel that it should always be possible to tether the discussion to foundational principles, which evolve manageably slowly, if at all.

8.1 Looking ahead

We have discussed above how the tension between strict regulation based on ML best practices and the practical demands of clinical trials on ML-based tools can be resolved. To support ML researchers, engineers and clinical development teams in finding case-by-case solutions, we suggest that this topic should be discussed more in ML forums. We encourage a move away from a competition-style approach to ML research that focuses on marginal performance gains on large, curated – and largely unrealistic datasets – towards reality-centric AI, i.e., focusing efforts on and rewarding impact of ML methods in the real world.

Furthermore, the example of DA methods in the late-stage trial calls for two parallel paths of action. Firstly, it is essential for ML engineers to repeat benchmarking exercises as in Guo et al [9] on datasets which reflect diverse types of clinical practice, e.g. oncology datasets of patient-specific treatment pathways mapped to survival outcomes or population-based cardiovascular datasets. A wide range of realistic benchmark experiments will allow the community to quantify both the deficit in the currently available armamentarium of ML methods with respect to clinical applications, as well as allow researchers to identify specific contexts where the current methods may offer real-world utility.

Secondly, the findings underscore the need to develop alternative approaches where deficits are identified. For example, for DA methods, there is an evolving consensus that the diversity of the large datasets used to train foundation models provide inductive biases, which may lead to improved extrapolation on a downstream prediction task [11]. At the same time, regulation on foundation models is nascent [1] [5] [12]. By engaging in dialogue with regulators and providing evidence for the benefits as well as the risks on specific ML methods, the ML community can positively contribute to the evolving ML regulatory guidance.

Finally, taking a step back, a long-term priority for the ML and healthcare community should be to collect large datasets that represent as many health contexts as possible and that will allow ML practitioners to build better models based on them.

9 Conclusion

In summary, we present a case study in applying ML tools to an oncology clinical trial to demonstrate the gap between ML best practices and regulatory guidance and the real-world challenges of implementing these standards. We focus on what we consider are the three domains critical for ML engineers developing tools in clinical trials: data representativeness, model interpretability and real-world usability. These areas raise practical regulatory concerns, often without a single generalisable solution.

Ultimately, ML for clinical development is a process designed to benefit patients. Our experience is that patients are open to the option of incorporating ML into their clinical care. It is imperative for the community and regulatory agencies to collaborate with a pragmatic mindset to achieve incremental and thoughtful progress for patients.

References

[1] European Parliament resolution of 14 June 2023 on the proposed regulation on Artificial Intelligence (AI Act). `https://www.europarl.europa.eu/doceo/document/TA-9-2023-0236_EN.html`, 2023. European Parliament Document TA-9-2023-0236.

[2] European Medicines Agency. Reflection paper on the use of artificial intelligence (AI) in the medicinal product lifecycle, 2023.

[3] N. ev, X. Glorot, J. W. Rae, M. Zielinski, H. Askham, A. Saraiva, A. Mottram, C. Meyer, S. Ravuri, I. Protsyuk, A. Connell, C. O. Hughes, A. Karthikesalingam, J. Cornebise, H. Montgomery, G. Rees, C. Laing, C. R. Baker, K. Peterson, R. Reeves, D. Hassabis, D. King, M. Suleyman, T. Back, C. Nielson, J. R. Ledsam, and S. Mohamed. A clinically applicable approach to continuous prediction of future acute kidney injury. *Nature*, 572(7767):116–119, Aug 2019.

[4] C. Loutradis, L. Pickup, J. P. Law, I. Dasgupta, J. N. Townend, P. Cockwell, A. Sharif, P. Sarafidis, and C. J. Ferro. Acute kidney injury is more common in men than women after accounting for socioeconomic status, ethnicity, alcohol intake and smoking history. *Biol Sex Differ*, 12(1):30, Apr 2021.

[5] National Institute For Standards in Technology. AI Risk Management Framework. `https://www.nist.gov/itl/ai-risk-management-framework`, 2023. Retrieved from National Institute for Standards in Technology.

[6] United States Congress. Algorithmic Accountability Act of 2023. `https://www.govinfo.gov/app/details/BILLS-118hr5628ih`, 2023. Retrieved from the U.S. Government Publishing Office.

[7] Fergus Imrie, Robert Davis, and Mihaela van der Schaar. Multiple stakeholders drive diverse interpretability requirements for machine learning in healthcare. *Nature Machine Intelligence*, 5(8):824–829, Aug 2023.

[8] Satyapriya Krishna, Tessa Han, Alex Gu, Javin Pombra, Shahin Jabbari, Steven Wu, and Himabindu Lakkaraju. The disagreement problem in explainable machine learning: A practitioner's perspective, 2022.

[9] Lin Lawrence Guo, Stephen R Pfohl, Jason Fries, Alistair E W Johnson, Jose Posada, Catherine Aftandilian, Nigam Shah, and Lillian Sung. Evaluation of domain generalization and adaptation on improving model robustness to temporal dataset shift in clinical medicine. *Sci. Rep.*, 12(1):2726, February 2022.

[10] Medley O'keefe Gatewood, Matthew Wemple, Sheryl Greco, Patricia A Kritek, and Raghu Durvasula. A quality improvement project to improve early sepsis care in the emergency department. *BMJ Qual. Saf.*, 24(12):787–795, December 2015.

[11] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2022.

[12] European Commission. Regulatory framework proposal on artificial intelligence. `https://digitalstrategy.ec.europa.eu/en/policies/regulatory-framework-ai`, 2021. Retrieved from European Commission, Digital Strategy.