M2 Internship and PhD Proposal @ Inria PreMeDICaL
—
Human-AI Interactions in Medical Contexts

This is a proposal for a 6-months M2 intership in the PreMeDICaL team at Inria Montpellier, under the supervision of Clément Berenfeld and Aurélien Bellet, with possible collaborations with Mathieu Even and Julie Josse. This internship is to be pursued into a 3-years PhD.

Keywords. Causal inference, human-AI interactions, performative learning, policy learning.

Contact. clement.berenfeld@inria.fr

Starting date. Spring 2026.

About the team. PreMeDICaL is a joint Inria-Inserm research unit based in Montpellier and led by Julie Josse. We develop statistical and machine learning methods for precision medicine, spanning topics such as causal inference, survival analysis, policy learning, federated learning, and private or fair ML. We are composed of both ML researchers and clinicians, as our work aims to cover the full path from theory to clinical deployment.

1. Motivations

In medical settings, where people's health is at stake, artificial intelligence (AI) systems are never deployed as autonomous decision-makers, and clinical decisions remain under the authority of human experts, who retain the final say over diagnostic and therapeutic actions. This interaction between AI-systems and human experts during deployment substantially increases the complexity of both evaluation and modeling. In this context, we would like to be able to answer these two questions: (i) Given an AI-model, can we assess whether its deployment will lead to improved patient outcomes compared to those achieved under the sole decisions of human experts? and (ii) Can we predict the post-deployment accuracy of a predictive AI-model that was not trained under deployment conditions?

To formalize this, let us denote by $A \in \mathcal{A}$ the clinical action (e.g. the choice of a treatment). The decision made in the presence of AI can be modeled as

$$A = h(X, f(X)),$$

where X denotes the patient's covariates, $f(X)$ is the AI prediction of a clinical outcome Y (e.g., a diagnosis), and h is some implicit random function. One would wish to chose f such that the action A leads on average to the best patient outcome Z (possibly different than Y, e.g., a survival time). Ideally, one wishes to compute

$$\mathrm{Val}(f) := \mathbb{E}_P\left[Z_{|A=h(X,f(X))}\right],$$

where $Z_{|A=h(X,f(X))}$ is the value of the outcome Z if, possibly contrary to the fact, the clinician had taken action $h(X,f(X))$ for patient X. Formal ways exist to define this quantity, see e.g. [9].

Evaluation challenges arise in settings where randomized controlled trials (RCTs) are not feasible and exposure to the model is purely observational. Data may alternate between periods with and without deployment, some clinicians may adopt the system while others don't, and some may consult the prediction only for certain patients while disregarding it for others. Such heterogeneous environments can complicate identification of the causal effect of integrating the model into clinical decision-making pipelines.

Another dimension concerns the alignment between model errors and clinician errors. Deployment value increases when the system succeeds precisely where clinicians fail. Overlaps of failures water down the usefulness of the model and can amplify risks, making the joint error structure a important component of evaluation.

Despite the causal structure of the problem, current AI evaluation pipelines are almost exclusively designed in standalone settings, where the AI models are assessed through metrics such as accuracy:

$$\mathrm{Acc}_P(f) = \mathbb{E}_P[\mathcal{L}(Y, f(X))],$$

for some loss function $\mathcal{L}$. Such an evaluation is often misleading and, paradoxically, models optimized for predictive accuracy may fail to improve or even degrade clinical outcomes once integrated into human workflows, see for instance [2, 11, 12]. The reasons behind this is that the deployment of prediction models inherently alters the data distribution, since the model's predictions shape the clinician behavior and, in turn, the patients outcomes. Once f is introduced into the decision pipeline, the observed data no longer follow the original distribution $P(X,Y)$, but a shifted one:

$$Q_{P,f}(X,Y) := P(Y \mid X, A = h(X, f(X))) \times P(X),$$

and this feedback mechanism can induce self-reinforcing biases and degrade the model over time. In the performative learning framework [10], this dynamic is expressed as $P_{t+1} = Q_{P_t,f_t}$ and $f_t \in \operatorname{argmin} \mathrm{Acc}_{P_t}$, with equilibrium reached at (P^*, f^*) when $P^* = Q_{P^*,f^*}$ and $f^* \in \operatorname{argmin} \mathrm{Acc}_{P^*}$. Clinical AI systems are never evaluated under such equilibrium conditions, which leaves the real-world accuracy of such models untested before deployment.

Developing a meaningful human-AI interaction frameworks requires therefore a careful formalization of the underlying causal mechanisms and of the deployment dynamics. This is what motivates this internship and PhD program.

2. Objectives of the internship and PhD

2.1 Theoretical developments

This internship and PhD ambition to build a formal framework for the design and evaluation of AI models deployed in medical contexts. The internship will focus on direction (1), centered on evaluating the deployment of AI-models. If you choose to continue with us as a PhD student, the project would then extend to directions (2) and (3) depending on your interests.

(1) Causal evaluation of AI-models. The estimation of $\mathrm{Val}(f)$ is an intricate problem that heavily depends on the context under which the AI-model is deployed. Each context, going in increasing complexity from idealized RCTs to fully observational environments, requires its own causal formulation of how AI outputs alter expert actions and how these altered actions propagate to patient outcomes. The aim is to construct, for each setting, an explicit causal model and corresponding identification results for $\mathrm{Val}(f)$, extending the type of analysis developed in the judicial framework of [3]. Clinical contexts relevant to this progression appear in Section 2.2.

(2) Policy learning of AI-exposure. When a model is not always accurate, or when it can be misleading, it might be useful to develop personalized rules to decide when to expose the clinicians to AI predictions. Building on policy learning [1], the task is to learn an exposure policy $\pi : \mathcal{X} \to \{0,1\}$ which decides when to display the prediction to the expert. Defining

$$f_\pi(X) = \begin{cases} f(X) & \pi(X) = 1 \\ \varnothing & \pi(X) = 0, \end{cases}$$

the optimal policy is the one that maximizes the causal value of the resulting human-AI system:

$$\pi^* \in \arg\max_\pi \mathrm{Val}(f_\pi).$$

The goal would be to find computable estimators for this optimal exposure map π^* and to derive its statistical properties.

(3) Performative learning and joint optimization. The previous objectives treat the AI-model f as fixed, but, realistically, one would wish to optimize these models for their deployment environment. In order to do so, it is natural to optimize for an objective that accounts for the deployment feedback described in Section 1:

$$f^* \in \arg\min_f \mathrm{Perf}(f) := E_{Q_{P,f}}[\mathcal{L}(Y, f(X))].$$

Such an optimization problem falls under the performative learning paradigm. So far, theoretical results only exist in simplified setting, see e.g. [4].

A logical extension is to jointly learn both the prediction model f and the exposure policy π. In some settings, e.g. for deferral learning, this joint optimization problem can take the form

$$f^*, \pi^* \in \arg\min_{f,\pi} \mathbb{E}_P[\pi(X)\mathcal{L}(Y, f(X))] + \mathbb{E}_P[(1-\pi(X))\mathcal{L}(Y, g(X))],$$

where $g(X)$ denotes a given human-expert prediction; in which case, only limited results exists for linear models, see e.g. [7]. Results in the context of this proposal are yet to be developed.

2.2 Use cases and applications

As part of your internship and PhD, you will have the opportunity to work with new and historical partners of PreMeDICaL. Here are three suggested use cases. All settings provide natural playgrounds for studying how AI recommendations interact with human decisions.

1. **Application to Traumabase.** Traumabase [6] is a national registry and research consortium focused on trauma care in France, which in particular targets the problem of pre-hospital triage, where paramedics must rapidly assess injury severity and decide on the appropriate hospital destination. They have developed a AI-model to estimate the probability of severe trauma from early observations [5], and are testing the integration of this model directly in ambulances in an ongoing RCT. This trial will test whether AI-assisted triage decisions improve survival of patients and reduce secondary transfers.

2. **Collaboration with CHU Montpellier.** The Pred'IC model is an AI prediction model developed by the firm KanopyMed which predicts the risk of rehospitalization in patients with heart failure. Its goal is to assist physicians in deciding the length of stays and follow-up cares, and thus aims to reduce congestion in cardiology wards. A RCT is being prepared by the CHU Montpellier to measure the effect of deploying this AI model on both patient outcomes and hospital throughput. Preliminary data exist regarding the performance of Pred'IC in pre-deployment settings.

3. **Collaboration with Doctolib.** A new potential partnership with Doctolib explores the deployment of an AI chatbot designed to support parents of newborns [8]. The chatbot provides guidance for early-life health questions and helps parents distinguish benign symptoms from those requiring medical attention. A RCT is in preparation to evaluate the impact of the chatbot usage on newborn health indicators and on behavioral outcomes such as avoidance of unnecessary emergency visits, and preliminary observational data are being collected among a few thousands of families.

References

[1] S. Athey and S. Wager. Policy learning with observational data. *Econometrica*, 89(1):133–161, 2021.

[2] D. R. Balcarcel, S. D. Mehta, C. G. Dixon, C. Z. Woods-Hill, E. C. Goligher, W. A. Van Amsterdam, and N. Yehya. Feedback loops in intensive care unit prognostic models: an under-recognised threat to clinical validity. *The Lancet Digital Health*, 7(8), 2025.

[3] E. Ben-Michael, D. J. Greiner, M. Huang, K. Imai, Z. Jiang, and S. Shin. Does AI help humans make better decisions? A statistical evaluation framework for experimental and observational studies. *arXiv preprint arXiv:2403.12108*, 2024.

[4] E. Cyffers, M. S. Pydi, J. Atif, and O. Cappé. Optimal classification under performative distribution shift. *Advances in Neural Information Processing Systems*, 37:68144–68160, 2024.

[5] T. Gauss, A. James, C. Colas, N. Delhaye, M. Holleville, B. Bijok, M. Werner, A. Meyer, V. Ramonda, E. Cesareo, et al. Comparison of machine learning and human prediction to identify trauma patients in need of hemorrhage control resuscitation (shockmatrix study): a prospective observational study. *The Lancet Regional Health–Europe*, 55, 2025.

[6] J.-D. Moyer, S. R. Hamada, J. Josse, O. Auliard, T. Gauss, T. Group, et al. Trauma reloaded: trauma registry in the era of data science. *Anaesthesia Critical Care & Pain Medicine*, 40(2):100827, 2021.

[7] H. Mozannar, H. Lang, D. Wei, P. Sattigeri, S. Das, and D. Sontag. Who should predict? exact algorithms for learning to defer to humans. In *International conference on artificial intelligence and statistics*, pages 10520–10545. PMLR, 2023.

[8] C. O'Brien and H. O'Reilly-Durand. La machine #48: Probabl's cause takes ai from lab to launchpad — plus: Doctolib's ai parenting assistant. *FrenchTech Journal*, October 2025. Accessed: 2025-11-13.

[9] J. Pearl. The do-calculus revisited. *arXiv preprint arXiv:1210.4852*, 2012.

[10] J. Perdomo, T. Zrnic, C. Mendler-Dünner, and M. Hardt. Performative prediction. In *International Conference on Machine Learning*, pages 7599–7609. PMLR, 2020.

[11] W. A. van Amsterdam, N. van Geloven, J. H. Krijthe, R. Ranganath, and G. Ciná. When accurate prediction models yield harmful self-fulfilling prophecies. *Patterns*, 6(4), 2025.

[12] N. van Geloven, R. H. Keogh, W. van Amsterdam, G. Cinà, J. H. Krijthe, N. Peek, K. Luijken, S. Magliacane, P. Morzywołek, T. van Ommen, et al. The risks of risk assessment: causal blind spots when using prediction models for treatment decisions. *Annals of internal medicine*, 178(9):1326–1333, 2025.