My vision is to develop **machine learning systems that are as rigorously tested and reliable as any medication or medical device**. My research lies at the intersection of *machine learning* and *causality*, with the goal of enabling reliable *prediction* and *decision-making*. In particular, I develop methods to **anticipate and correct unreliable behaviors** in machine learning systems before they cause harm.

This is an urgent problem. Machine learning is deployed today in high-risk domains like healthcare, but can fail in unanticipated ways. As a cautionary example, consider the Epic Sepsis Model, a proprietary prediction model used in hundreds of hospitals to detect a deadly condition (sepsis) that kills hundreds of thousands of patients per year. A recent study found that this model had far lower performance in practice (0.63 AUC) than originally found by the company (0.76–0.83 AUC), and an alarmingly high rate of false positives [17]. In part, this under-performance was due to an over-reliance on correlations in training data that were not present during deployment, such as indicators that clinicians had already begun treatment of the condition [10].

This example illustrates that models for *prediction* (e.g., for detecting sepsis) can perform poorly when correlations change, due to changes in patient populations or clinical practice. Meanwhile, models used to inform *decision-making* (e.g., for making treatment recommendations) can be misled by confounding and other biases in historical data. I develop tools to anticipate and correct failures in both types of models, by taking a **causal perspective**:

- **Reliable causal inference and policy evaluation**: Evaluation of models for *decision-making* requires strong causal assumptions when using observational data (vs. data from a clinical trial), and can be unreliable if those assumptions are violated. To help domain experts assess the credibility of treatment recommendations derived from observational data, I have developed methods to uncover unrealistic counterfactual claims [12, 5], and identify poorly represented sub-populations [13]. I have also developed methods to incorporate limited clinical trial data [4, 15] to improve credibility.
- **Robust, reliable prediction via causal knowledge**: For *prediction* models, causality is a useful lens for reasoning about how plausible changes in distribution will impact future model performance. In linear settings, I have developed methods for learning predictors with provably robust performance across changes in factors that are not directly observed (e.g., differences in socioeconomic status of patients) [14]. In more general settings, I have developed new ways for domain experts to express their causal intuition about plausible changes (e.g., changes in clinical practice), evaluate the worst-case performance of models under those changes, and discover changes that result in poor performance [16].

I approach these problems from a methodological perspective, publishing in machine learning venues like NeurIPS [16, 4, 8], ICML [12, 14], AISTATS [13], and KDD [1]. My methodological research is **inspired by my collaborative work with clinicians on healthcare applications**, including the methodological development of a model for antibiotic recommendations published in *Science Translational Medicine* [6]. I have also taken classes at Harvard Medical School on pathology and physiology, to improve my ability to collaborate with clinicians.

# Prior Research

## Reliable causal inference and policy evaluation

Learning and evaluating new treatment policies from retrospective (or "observational") data requires causal assumptions, such as full observation of confounding factors. My research has produced new tools to help "sanity check" the retrospective analysis of treatment policies, revealing failures of these causal assumptions and informing the design of more credible decision-making systems.

**Validating causal models by inspecting counterfactual claims** [12, 5, 11]: How should clinicians evaluate the claim that a new treatment policy will improve outcomes? In [12] (ICML 2019), I developed a technique to help clinicians validate probabilistic causal models used in sequential decision-making problems (e.g., in model-based reinforcement learning for sepsis treatment). My approach **decomposes aggregate claims** made by such models (e.g., "80% of hypothetical future patients would survive under the new policy") into **counterfactual claims** on specific historical patients (e.g., "patient X would have survived under the new policy"). To uncover flaws in the clinical reasoning of the model, these counterfactuals can be **reviewed by clinicians** alongside the full medical record for those patients. In [5] (AMIA 2021) and [11] (M.S. Thesis), I used this approach to **uncover flaws in a highly-cited paper**, the "AI Clinician" [7]. Alongside an infectious disease clinician from Mass General Brigham, I reviewed patients who died in reality, but who the model claimed *would have lived* under its recommended policy. Implausible counterfactual claims were often clearly attributable to confounding factors not included in the model (e.g., terminal cancer), but present in clinical notes. Such insights are directly relevant for improving model design (e.g., by extracting additional features to include as potential confounders).

The main technical innovation was to derive counterfactual claims from any existing model of discrete dynamics. This presented a problem: Counterfactual simulation requires specification of a structural causal model (SCM), but there are many SCMs that are consistent with the original model. Here, any SCM will produce a valid decomposition, but some decompositions are more interpretable than others. To this end, I introduced a condition called "counterfactual stability" that imposes common-sense restrictions on counterfactuals, and introduced a novel SCM that satisfies this condition. For instance, if we observe an increase in blood pressure in the absence of treatment, then we should also see a counterfactual increase in blood pressure if given a blood-pressure-increasing medication. This research **spurred interest from other research groups, leading to follow-up research** on alternative counterfactual restrictions [9] and use cases for counterfactual simulation [2].

**Identifying poorly-represented sub-populations** [13]: How can clinicians tell if the conclusions of a causal analysis apply to a particular patient? A necessary (and testable) condition is that similar patients were observed receiving both the treatment and control. In [13] (AISTATS 2020), I gave an algorithm (OverRule) for creating interpretable descriptions of the well-represented population, which could then be published alongside a retrospective study. This method was developed with a clinical collaborator from Beth Israel Deaconess Medical Center, inspired by applications in estimating the effect of post-surgical opioid prescriptions on future misuse, using health insurance claims data. The resulting output was evaluated in user studies with a small group of clinicians, and found to represent plausible clinical patterns. For instance, large opioid doses are rarely prescribed for C-section surgeries, and hence we cannot reliably infer causal effects of large vs. small doses in this population.

**Incorporating experimental data to improve credibility** [4, 15]: Experimental data (i.e., from a clinical trial) is often small-scale and narrow in scope. For instance, Phase 3 clinical trials for COVID vaccines did not originally include pregnant women [3]. To assess causal effects in these unrepresented populations, we often turn to observational data. In [4] (NeurIPS 2022) I demonstrated that the experimental data is still useful, despite not covering the population of interest. In particular, I developed a method that can be applied when multiple observational studies cover both the population of interest and the experimental subpopulation. The core idea is to use the experimental data to test for potential bias and then conservatively aggregate estimates across observational studies that pass this test. This is a form of meta-analysis (the analysis of multiple studies) that comes with guarantees under weaker assumptions. Instead of requiring that all studies are unbiased (e.g., free of confounding), this approach only requires that at least one observational study is unbiased. In ongoing work [15], I have developed a simple approach for combining experimental and observational data to estimate causal effects, when populations do overlap. The resulting estimate is consistent and has bounded worst-case performance regardless of the unknown bias (e.g., due to confounding) in the observational data.

## Robust, reliable prediction via causal knowledge

Predictive models can fail due to unreliable correlations that change across hospitals or patient populations. My research has produced techniques for **anticipating and avoiding these failures in advance.** I have focused on the **proactive** setting, where we only have access to data from the training distribution. In this setting, partial causal knowledge allows us to reason about performance of predictive models in unseen future scenarios.

Building reliable but effective models requires trade-offs. For instance, suppose a model for diagnosis depends on laboratory tests as a feature. Changes in laboratory testing policies could impact the correlation between these features and disease. A drastic approach to learning reliable models would discard lab-related features altogether, at the cost of lower predictive performance. The methods I develop allow for more principled trade-offs between reliability and effectiveness, by considering worst-case performance under plausible changes.

**Learning (linear) predictors with domain-specific reliability guarantees** [14]: In healthcare, many important variables are not directly observed, like social determinants of health (e.g., socioeconomic status). How should we train models proactively to have reliable performance when the distribution of these variables changes? In [14] (ICML 2021), I derived predictors that have optimal worst-case performance, in simplified linear settings, over a user-selected "robustness set" of plausible changes (e.g., a range of possible differences in average income between the training and deployment hospitals). Larger robustness sets favor models with invariant (but perhaps poor) performance, while additional knowledge (in the form of constraints on the set) allow for improved trade-offs between reliability and effectiveness. I demonstrated that these optimal predictors can be learned even when the changes occur in unobserved variables that are only observed at training time via multiple noisy proxies (e.g., self-reported data on income). Modeling the impact of these changes requires only partial causal knowledge, such as knowing that the unobserved variables are not causally affected by observed variables. This work is also relevant for **adapting models to new settings:** Given the mean and covariance of a single proxy in the test distribution, one can learn a model (using the training data) with optimal performance on the test domain.

**Evaluating performance under plausible worst-case scenarios** [16]: How should we proactively assess the reliability of a predictive model, given only data from the training distribution? In [16] (NeurIPS 2022), I gave a method for evaluating the worst-case performance of predictive models under changes in user-specified factors (e.g., laboratory testing). In contrast to my prior work [14], which considered linear settings, this approach applies to settings like computer vision with deep models, assuming only the availability of auxiliary features and partial knowledge of causal structure. The method allows users to both *discover* model vulnerabilities by searching over a large set of changes, and *precisely quantify accuracy-reliability trade-offs* between different models. These insights can directly inform model design. For instance, in the laboratory testing example above, a model that uses laboratory tests may still have better worst-case performance (under realistic changes in testing policies) than a model that discards these features, enabling experts to confidently choose to keep the features.

On a technical level, this approach involved two innovations. First, I developed a new approach to defining changes in distribution (or "shifts"), via parametric perturbations to the marginal or conditional distributions of a subset of variables. In contrast to work that considers arbitrary subpopulations or distributions in an $f$-divergence ball, this approach allows domain experts to further constrain the set of plausible shifts (e.g., restricting to a uniform increase or decrease in the availability of laboratory testing) and in some cases directly interpret the worst-case shift to build intuition for model vulnerabilities. Second, I introduced an empirical objective based on a second-order Taylor approximation (in the space of distributions) that yields a tractable optimization problem for finding the worst-case shift. Compared to a re-weighting-based approach, this approach avoids the need for solving a non-convex, high-variance optimization problem, and empirically uncovers more impactful shifts.

## Future Research

Machine learning systems in high-risk settings will always need to adapt to new situations. Healthcare itself is constantly evolving due to changes in clinical practice, disease patterns, data collection, and diagnostic technology. I aim to make it safe and easy for clinicians and data scientists to adapt machine learning models for new contexts, while ensuring that the models remain rigorously tested and reliable. More broadly, my research is applicable in other high-stakes settings where machine learning is deployed (like criminal justice and consumer lending).

**Building credible decision-making models using limited experimental data**: Experimental data (e.g., from randomized trials) is free of confounding, but is often smaller in scale and not representative of the whole population. This data can serve as a useful tool for benchmarking observational studies, to see if they recover similar estimates of treatment effects. But how should this data guide model development, when estimates do not align? My prior and ongoing work [4, 15] demonstrates that there is no "free lunch": We cannot guarantee improvement upon experimental estimates without some causal assumptions, but I am interested in exploring weaker causal assumptions (e.g., limited confounding) under which such improvement is possible.

**Automating the search for unreliable correlations in prediction**: My prior work in reliable prediction has focused on worst-case evaluation [16] and optimization [14] under plausible changes in distribution, but it requires specifying a set of plausible changes in advance. How should we uncover unreliable correlations that we are not already aware of, and gauge the impact on model performance in practice? Here, I believe there is an opportunity to leverage rich structured or unstructured data on clinical context (e.g., the full medical record for an X-ray, in medical imaging applications), to discover and quantify the impact of potentially spurious signals. In addition, I am interested in developing methods for attribution: If predictive performance (across time, or across locations) drops off, can we attribute this to specific, human-interpretable changes in distribution?

**Controlling how prediction models adapt to new contexts**: Not all hospitals have sufficient data to train a new model from scratch on their own population. However, adapting existing models (trained on a broader range of hospitals) presents unique challenges in a healthcare context. First, we may wish to adapt to certain changes (e.g., a different mix of patients) without adapting to others (e.g., spurious correlations between X-ray scanner brand and disease). Moreover, hospitals may differ in the reliability of structured features (e.g., diagnosis codes), the availability of laboratory tests, and so on. One promising direction is to infer the relevant changes from limited data and some knowledge of causal structure, conceptually similar to my work on adaptation in linear settings [14]. However, applying this idea in general settings raises challenges of identification: For instance, we may observe fewer tests due to unobserved changes in patient populations, or because fewer tests are available.

**Deploying and monitoring real-world systems**: Following my work on learning effective antibiotic treatment policies [6], my clinical collaborators are planning to prospectively evaluate a decision-support system across the Mass General Brigham and NYU-Langone hospital systems. I intend to continue pursuing interdisciplinary collaborations in high-stakes settings, to ground my methodological research in the practical challenges of deploying and monitoring machine learning systems in the real world.

**Conclusion** Machine learning has incredible potential to change the way that we care for patients. However, deploying machine learning with confidence requires better tools for model design, pre-deployment evaluation, adaptation to new contexts, and ongoing monitoring. Ultimately, my goal is to see machine learning become part of the standard of care, as reliable and trustworthy as any drug or diagnostic test we use today. I believe that research along the lines described here will be instrumental in bringing that future to pass.

# References

[1] Boominathan, S., **Oberst**, M., Zhou, H., Kanjilal, S., and Sontag, D. Treatment policy learning in multiobjective settings with fully observed outcomes. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '20, pp. 1937–1947, New York, NY, USA, August 2020. ACM.

[2] Corvelo Benz, N. L. and Gomez Rodriguez, M. Counterfactual inference of second opinions. In *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*, 2022.

[3] Dagan, N., Barda, N., Biron-Shental, T., Makov-Assif, M., Key, C., Kohane, I. S., Hernán, M. A., Lipsitch, M., Hernandez-Diaz, S., Reis, B. Y., and Balicer, R. D. Effectiveness of the BNT162b2 mRNA COVID-19 vaccine in pregnancy. *Nature medicine*, 27(10):1693–1695, October 2021.

[4] Hussain*, Z., **Oberst***, M., Shih*, M.-C., and Sontag, D. Falsification before extrapolation in causal effect estimation. In *Advances in Neural Information Processing Systems*, September 2022.

[5] Ji*, C. X., **Oberst***, M., Kanjilal, S., and Sontag, D. Trajectory inspection: A method for iterative clinician-driven design of reinforcement learning studies. In *AMIA Virtual Informatics Summit*, 2021.

[6] Kanjilal, S., **Oberst**, M., Boominathan, S., Zhou, H., Hooper, D. C., and Sontag, D. A decision algorithm to promote outpatient antimicrobial stewardship for uncomplicated urinary tract infection. *Science translational medicine*, 12(568), November 2020.

[7] Komorowski, M., Celi, L. A., Badawi, O., Gordon, A. C., and Faisal, A. A. The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. *Nature medicine*, 24(11):1716–1720, October 2018.

[8] Lim*, J., Ji*, C., **Oberst***, M., Blecker, S., Horwitz, L., and Sontag, D. Finding regions of heterogeneity in decision-making via expected conditional covariance. In *Advances in Neural Information Processing Systems*, 2021.

[9] Lorberbom, G., Johnson, D. D., Maddison, C. J., Tarlow, D., and Hazan, T. Learning generalized gumbel-max causal mechanisms. In *Advances in Neural Information Processing Systems*, 2021.

[10] Ross, C. Epic's sepsis algorithm is going off the rails in the real world. the use of these variables may explain why. `https://www.statnews.com/2021/09/27/epic-sepsis-algorithm-antibiotics-model`, September 2021. Accessed: 2022-10-19.

[11] **Oberst**, M. Counterfactual policy introspection using structural causal models. Master's thesis, Massachusetts Institute of Technology, 2019.

[12] **Oberst**, M. and Sontag, D. Counterfactual off-policy evaluation with gumbel-max structural causal models. In *Proceedings of the 36th International Conference on Machine Learning*, 2019.

[13] **Oberst***, M., Johansson*, F., Wei*, D., Gao, T., Brat, G., Sontag, D., and Varshney, K. Characterization of overlap in observational studies. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, 2020.

[14] **Oberst**, M., Thams, N., Peters, J., and Sontag, D. Regularizing towards causal invariance: Linear models with proxies. In *Proceedings of the 38th International Conference on Machine Learning*, 2021.

[15] **Oberst**, M., D'Amour, A., Chen, M., Wang, Y., Sontag, D., and Yadlowsky, S. Bias-robust integration of observational and experimental estimators. *arXiv preprint (2205.10467)*, May 2022.

[16] Thams*, N., **Oberst***, M., and Sontag, D. Evaluating robustness to dataset shift via parametric robustness sets. In *Advances in Neural Information Processing Systems*, 2022.

[17] Wong, A., Otles, E., Donnelly, J. P., Krumm, A., McCullough, J., DeTroyer-Cooley, O., Pestrue, J., Phillips, M., Konye, J., Penoza, C., Ghous, M., and Singh, K. External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. *JAMA internal medicine*, 181(8):1065–1070, August 2021.