Research Statement

Roy Adams (radams47@jhu.edu)

Observational health data – collected via mobile sensors and electronic health records – presents an unprecedented opportunity to study health behaviors and chronic conditions, find sources of heterogeneity in patient populations, and target care to at-risk patients [1] [2] [3] [4] [5]. Taking advantage of this opportunity will require solving a variety of challenging computational and data analytical problems. *My goal as a machine learning researcher is to develop models, algorithms, and tools that leverage this data to provide insight into complex health conditions and improve the safety, efficiency, and effectiveness of patient care.* In particular, my research focuses on developing machine learning and statistical methods to answer three core questions:

1. How do we make reliable inferences from health data affected by measurement error, confounding, and societal biases?
2. How do we identify which patients are most at-risk and which treatment is best for each patient?
3. How do we determine whether a machine learning model is safe and ready to deploy in a healthcare environment?

While primarily motivated by problems in healthcare, my work addresses core methodological questions in probabilistic modeling, causal inference, and model evaluation with implications to a broad range of machine learning applications. My approach to research begins by identifying opportunities to leverage health data through collaboration with clinicians, behavioral scientists, and public health researchers and, as such, interdisciplinary collaboration is central to my work. In the remainder, I will describe my past, ongoing, and future research grouped around the three questions raised above.

Making reliable inferences from health data

Data arising from real-world healthcare environments is affected by measurement error and missing data, observed and unobserved confounding, and a variety of societal biases. A major goal of my work is to develop methods for drawing reliable, unbiased conclusions when using such data. This means developing technologies that provide reliable measurements of health and accounting for potential measurement error and bias when such measurements are unavailable.

Activity detection from wearable sensors (past work):

Health behaviors such as smoking, eating, drug use, and exercise have traditionally been studied using self-report data; however, self-report has well-known limitations including data sparsity, recall bias, and high burden on study participants. In contrast, wearable sensors allow us to observe participants at high frequency, in non-clinical settings, and without relying on patient self-report. In papers published in ICML [6], IEEE Wireless Health [7], and IMWUT [8], I led development of a family of segmentation-based graphical models for detecting common health behaviors from wearable sensor data. We used these models to detect smoking, eating, sleeping, conversation, and electrocardiogram morphology. By modeling regular episodic structure in such behaviors, we were able to learn these models from relatively limited amounts of labeled data.

Measurement error in observational health data (ongoing work):

When reliable measurements are not available, it is critical that we account for systematic measurement error to avoid drawing biased inferences. Unfortunately, observational health data frequently violates the assumptions of classical measurement error analysis and we lack methods to incorporate many measurement error assumptions that commonly occur in healthcare settings. My work has addressed these

challenges in three ways. First, in papers published in AISTATs [9] and UAI [10], I led development of learning methods for activity detection models when activity labels are temporally imprecise or missing. Second, in work published in ICML [11], I developed a method for estimating the causal effect of a binary exposure subject to *underreporting* – a common type of measurement error in survey data – and used this method to estimate the effect of maternal opioid use on later childhood health. Finally, in work currently in submission as AISTATs, we showed how partial identification bounds for a parameter of interest can be computed in a large class of commonly occurring discrete variable measurement error models. This class includes common models – such as the instrumental variable model with measurement error on the outcome variable – allowing inference in these models without relying on unrealistic assumptions.

Future work:

In the short term, I will continue to investigate the many common and important models which are known not to be fully identifiable, but for which no known partial identification bounds exist. I propose to continue work on partial identification in measurement error models to (1) identify new classes of models that are amenable to methods from my previous work and (2) derive alternative strategies for models that are not amenable to these methods. These strategies may include various approximations that produce non-sharp or approximate bounds.

In the long term, I propose to study the effect of measurement error on predictive models in health care and develop methods to account for additional uncertainty or bias that such errors may cause. Despite the near ubiquity of measurement error in observational health data, relatively little is known about how systematic measurement error can impact machine learning-based predictive models. If unaccounted for, measurement error may hurt a model's performance or make a model overly confident in an incorrect prediction. This is especially concerning if, for example, measurement error rates are higher in a vulnerable subpopulation, such as when a particular condition is under-diagnosed or coded differently in that subpopulation. This will entail both empirical study of how models perform in settings with different types of measurement error, as well as methodological work on incorporating measurement error assumptions into the model learning process and quantifying additional uncertainty caused by measurement error.

Identifying at-risk patients and targeting care

Delivering the most effective and efficient care possible requires identifying (1) which patients will benefit the most from treatment and (2) which treatment is best for each patient. Heterogeneity in patient populations can make these tasks difficult as patients at similar risk levels may present differently and may respond differently to the same treatment. Observational health data can allow us to identify markers of heterogeneity in a population, signal healthcare providers to at-risk patients they may not be aware of, and make personalized treatment recommendations.

Targeted health messaging (past work):

Behavioral interventions, such as those used to help patients quit smoking, may be sensitive to heterogeneity in patients' social or cultural contexts. In work published in AMIA [12], RecSys [13], and JMIR [14], I led development of a recommendation system that personalized the content of motivational messages sent to patients trying to quit smoking. This system was deployed as part of an online smoking cessation portal and, in a randomized trial, we found that this type of targeting increased the proportion of messages that patients felt were relevant.

Future work:

As part of my long-term research program, I will continue to identify areas where treatment guidelines lack the necessary flexibility or specificity to be maximally effective and, in collaboration with healthcare providers, design tools that can improve the delivery of care. One application of particular interest is identifying which patients are responding to antibiotics in critical care settings. Clinicians treating infection in such settings must constantly balance the benefits of early administration of antibiotics against the risks of overtreatment, often relying on trial and error. Detecting, as early as possible, if a patient is responding to a particular antibiotic will allow clinicians to identify the right antibiotic sooner and reduce the time patients spend on unnecessary antibiotics.

Deploying machine learning systems in healthcare settings

Despite a recent surge in the application of machine learning models to health data, there is still much we do not know about deploying such models to real healthcare settings. Standard methods for evaluating machine learning models do not tell us how safe a model is to use, the effect a model will have on patient outcomes, or how a model will impact health disparities. Further, the effectiveness of a machine learning model depends not just on the predictive performance of the model, but on how useful and trustworthy clinicians find the outputs of the model. In ongoing and future work, I propose to build a standard by which we can judge whether a machine learning system is ready to deploy in a healthcare setting.

Evaluating the effect a model has on health outcomes (ongoing work):

The primary goal of model evaluation in healthcare settings is to understand how deploying a model will impact the health outcomes of patients. However, in many settings, predictive performance cannot easily be translated into the causal effect a model will have when deployed. Our approach to this problem is to map the outputs of a machine learning model to a treatment policy, and then to estimate average patient outcomes under that policy using causal estimation methods. In ongoing work, I have applied this approach to estimate the effect of a sepsis early warning system on patient outcomes. This approach necessitates the application of causal estimation methods to continuous-time health record data; however, the behavior of standard causal estimators on this type of data is not well understood. In work in submission at Biostatistics, I led an investigation of the performance of discrete-time causal estimation methods on continuous-time medical data and provided practical guidance for how to avoid potential sources of estimation bias incurred by discretizing the data.

Evaluating model generalization (ongoing work):

Classical model evaluation methods fail to capture how the model will *generalize* new environments that differ from the test data. Methods that do measure generalization typically require gathering representative datasets from multiple environments, which is often infeasible. In work in submission at AISTATs, we developed an approach for evaluating model transportability using data from only a single environment. This approach uses the test data to define an *uncertainty set* of possible environments, and we proposed a statistically efficient estimator for the *worst-case* model performance across all environments in this set.

Future work:

In the short term, I propose to continue developing a methodology for prospectively evaluating the causal effect of deploying a machine learning model. This will involve formalizing the process by which we map a model's predictions to a treatment policy, including possible modeling of how healthcare providers respond to those predictions. Additionally, I will continue to study best practices for applying existing

causal estimation methods to complex health data and to develop new estimators where methodological gaps exist.

In the long term, I propose to study how healthcare providers interact with machine learning models once they are deployed. In order for a model to improve patient outcomes, clinicians must trust the outputs of the model. Despite this, very little is known about how healthcare providers develop trust in a model and how that trust may evolve over time. Open questions include: How do different types of prediction errors affect provider trust? How do provider expertise and experience impact a provider's willingness to trust a model? And, how do provider workload and stress impact willingness to rely on a model? Answering these questions will require a combination of controlled studies using simulated systems and direct study of how healthcare providers interact with existing deployed models.

References

[1] A. O'Connor, "A.I. Is Likely to Transform Medicine," New York Times, 12 March 2019. https://www.nytimes.com/2019/03/11/well/live/how-artificial-intelligence-could-transform-medicine.html.

[2] "Innovations In: AI and digital health," Scientific American, 14 January 2020. https://www.scientificamerican.com/custom-media/innovations-in-ai-and-digital-health/.

[3] A. Almendrala, "7 Ways Apple's New Software Could Change Medical Research For The Better," Huffington Post, 10 March 2015. https://www.huffpost.com/entry/apple-researchkit-health_n_6840484.

[4] E. Woykearchive, "These Wearables Detect Health Issues Before They Happen," MIT Technology Review, 30 November 2016. https://www.technologyreview.com/2016/11/30/6210/these-wearables-detect-health-issues-before-they-happen/.

[5] S. Saria, "A $3 trillion challenge to computational scientists: transforming healthcare delivery," *IEEE Intelligent Systems,* vol. 29, no. 4, pp. 82-87, 2014.

[6] **R. Adams**, N. Saleheen, E. Thomaz, A. Parate, S. Kumar and B. Marlin, "Hierarchical Span-Based Conditional Random Fields for Labeling and Segmenting Events in Wearable Sensor Data Streams," in *International Conference on Machine Learning*, 2016.

[7] T. Nguyen, **R. Adams**, A. Natarajan and B. Marlin, "Parsing Wireless Electrocardiogram Signals with Context Free Grammar Conditional Random Fields," in *IEEE Wireless Health*, 2016.

[8] R. Bari, **R. Adams**, M. M. Rahman, M. B. Parsons, E. Buder and S. Kumar, "rConverse: Moment by Moment Conversation Detection Using a Mobile Respiration Sensor," *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies,* vol. 2, no. 1, pp. 1-27, 2018.

[9] **R. Adams** and B. Marlin, "Learning Time Series Detection Models from Temporally Imprecise Labels," in *International Conference on Artificial Intelligence and Statistics*, 2017.

[10] **R. Adams** and B. Marlin, "Learning Time Series Segmentation Models from Temporally Imprecise Labels," in *Uncertainty in Artificial Intelligence*, 2018.

[11] **R. Adams**, Y. Ji, X. Wang and S. Saria, "Learning Models from Data with Measurement Error: Tackling Underreporting," in *International Conference on Machine Learning*, 2019.

[12] B. Marlin, **R. Adams**, R. Sadasivam and T. Houston, "Towards Collaborative Filtering Recommender Systems for Tailored Health Communications," in *American Medical Informatics Association Annual Symmposium*, 2013.

[13] **R. Adams**, R. Sadasivam, K. Balakrishnan, R. Kinney, T. Houston and B. Marlin, "PERSPeCT: Collaborative Filtering for Tailored Health Communications," in *ACM Conference on Recommender Systems*, 2014.

[14] R. Sadasivam, E. Borglund, **R. Adams**, B. Marlin and T. Houston, "Impact of a collective intelligence tailored messaging system on smoking cessation: the Perspect randomized experiment," *Journal of Medical Internet Research,* vol. 18, no. 11, p. e285, 2016.