Implementing AI in healthcare

Vector-SickKids Health AI Deployment Symposium, Toronto, Ontario, Canada
Event Date: October 30, 2019
Published Date: March 24, 2020

Erik Drysdale[1*], Elham Dolatabadi[2,3], Corey Chivers[4], Vincent Liu[5], Such Saria[6], Mark Sendak[7], Jenna Wiens[8], Michael Brudno[1,2,3], Amelia Hoyt[1], Mjaye Mazwi[1], Muhammad Mamdani[2,3,9], Devin Singh[1], Vanessa Allen[10], Carolyn McGregor[11], Heather Ross[12], Antonio Szeto[13], Amol Anand Verma[2,8], Bo Wang[2,11,13], P. Alison Paprica[2,3], Anna Goldenberg[1,2,3]

[1] The Hospital for Sick Children
[2] University of Toronto
[3] The Vector Institute
[4] The University of Pennsylvania Health System
[5] Kaiser Permanente Northern California Division of Research
[6] Johns Hopkins University
[7] Duke Institute for Health Innovation
[8] University of Michigan in Ann Arbor
[9] St. Michael's Hospital
[10] Public Health Ontario
[11] Ontario Tech University
[12] University Health Network
[13] University of Waterloo

*Corresponding author: erik.drysdale@sickkids.ca

https://vectorinstitute.ai/health-research/

Executive Summary

Advances in artificial intelligence (AI), and its subfield machine learning (ML), can be seen in almost every domain of life, including cutting-edge health research.[1,2] However, only a tiny fraction of health AI/ML systems described in research papers makes its way into clinical practice. To help address this issue the Hospital for Sick Children (SickKids) and the Vector Institute for Artificial Intelligence (Vector) organized the Vector-SickKids Health AI Deployment Symposium on October 30, 2019 attended by 166 clinicians, computer scientists, policy makers, and healthcare administrators. The aim was to showcase real-world examples of AI moving from the research lab to the clinic. Speakers came from a variety of Canadian and US institutions including St. Michael's Hospital, the University Health Network, the University of Waterloo, Public Health Ontario, Ontario Tech University, the University of Michigan, Northern California Kaiser Permanente, Johns Hopkins University, University of Pennsylvania, and Duke University. The successes and challenges that each project experienced provided valuable insights into the new and evolving field of AI for health. Each speaker was asked to prepare a structured presentation which touched upon the following topics:

- Prerequisites for deployment (such as data access)
- AI Applications
- Evaluation procedures
- Visualization strategies
- Team building
- Ethical considerations
- Deployment pipeline
- Lessons learned

The focus was on identifying concrete "do's and don'ts" for deploying ML into healthcare.

The deployment strategies for ML in healthcare are currently ad hoc for most applications. A lack of defined rules and best practices leads to provisional solutions that may be suboptimal. Through a better knowledge of real-world implementations several common themes surfaced. These examples are a first step in understanding what is required to define pathways for AI/ML deployment. Academic researchers in the field of health AI/ML generally focus on factors required for successful *data science* ranging from statistical analysis and ML algorithms to database access and institutional review board approval. *Implementation science* is an equally important discipline that is needed to bring ML tools to the bedside.[3] This field of research is by no means new, and has been studied extensively in the context of translating medical research into clinical practice.[4] In a related vein, understanding how and why institutions adopt and maintain technologies is also an important component of any healthcare technology and change management more broadly.[5,6] Implementation science questions in ML for healthcare include 'What are the operational components needed to maintain and monitor a system?' and 'How will

feedback be enabled and incorporated into practice?'. These questions need to be answered for the appropriate design and successful deployment of AI/ML models. During model development, in order to choose algorithms, features, and evaluation metrics that will lead to robust and institutionally appropriate systems, research teams need to take the entire pipeline of project development into view.

Three key interrelated themes were raised throughout the symposium: contextualization, life-cycle planning, and stakeholder involvement.

Contextualization: AI/ML tools that are deployed must be contextualized in an existing workflow. Technology necessarily operates within existing norms and practices--especially in healthcare[7]. For example, in the case of a tool that is meant to detect sepsis early, a researcher must be engaged with how the disease is currently being detected in the hospital. If the institution does a good job at detecting all patients who develop sepsis then perhaps the algorithm can help to expedite this detection process. In contrast, if some patients have a diagnosis that is being systematically missed, evaluation should focus on these challenging patients. By contextualizing algorithms in an existing workflow, researchers will ensure they are providing appropriate solutions to existing problems rather than ones that are of less interest to the end users. Tools designed only for *in-silico* validation will inevitably struggle to obtain uptake by physicians and nurses and limit project success. Understanding clinical context also allows algorithm evaluation to match the real-time environment. Especially for early warning systems, having IT systems which record accurate time-stamps is essential to establishing outcome-independent time periods in hospital settings.

Life-cycle Planning: All speakers at the symposium identified institution-specific project cycle management schema that are being used to build systems in the fastest possible time with the largest clinical benefit. These frameworks generally include three components: the research/scientific stage, the technical/implementation stage, and the operational/maintenance stage. The overwhelming number of papers being published in top-tier ML and/or healthcare journals are focused on the research/scientific stage where attention is focused on the details of dataset features and *in-silico* algorithm performance. At the technical/implementation stage, questions regarding how fast data can be extracted from a hospital's EHR system and how the system will be evaluated in real-time need to be addressed. The use of "silent period" occurs in the technical implementation stage when an algorithm is first embedded into an institution's IT system and makes real-time predictions without communicating them or impacting clinical care (see Appendix A2 for more discussion of a model's silent period). Further, while existing statistical procedures to evaluate model performance can be used for these trials, there are additional considerations when the project moves to the operational/maintenance stage around how the algorithm is to be updated and how it will respond to changes in the underlying data stream (such as new data fields). Having feedback for AI projects is essential for ensuring sustainability and trust. Commercial AI tools are constantly being updated with new data, edge cases, and examples of failures. Responsible AI/ML requires that there is accountability and clarity about who is responsible for ongoing adjustments to the algorithm, assessing

human-computer interactions, resource procurement, and timely delivery to the front end. On an institutional level, this means having a strong mandate that can help to sustain model development and pull in the necessary resources. These adjustments and accountability must be linked to broader considerations around fairness, including issues of racial bias and unintended consequences for vulnerable groups.

Stakeholder involvement: No algorithm is an island unto itself. Every successful project presented at the symposium was enabled by the intersection of operational and research leadership along with a variety of clinical stakeholders. This is essential to address important risks such as alarm fatigue, i.e., the possibility that AI system alerts are increasingly ignored by clinicians. Interestingly, a common theme among presenters was that during initial deployment (the silent period), the AI/ML prediction would be sent to someone other than the physician or nurse who assesses and validates the signal in context before passing it on for action. As health AI/ML research matures, we hope to see a deployment mindset affect how research is carried out and discussed. Commonly reported metrics such as the Area Under the Receiver Operator Characteristic (AUROC), may be useful for understanding the level of signal in the data, but need to be accompanied by actually calibrated metrics such as the positive predictive value if they are going to resonate with and be meaningful for clinicians. AI/ML researchers also need to consider how false positives and negatives are to be balanced, and the risks/costs associated with specific thresholds. Deeper involvement of stakeholders can take many forms. For example, rather than having discussions around how an algorithm could be implemented in the “future direction” sections of academic papers, they could be moved directly into the main body of research signalling that they are considered from the beginning of model development.

AI/ML tools in healthcare can help to improve patient outcomes, reduce costs, and improve the workplace experience of healthcare practitioners, but they have to be taken out of research and into practice in a responsible way. Despite the thousands of research articles about AI/ML that have the potential to improve health and healthcare, information about how to responsibly deploy health AI/ML models is hard to find[8]. By providing structured information about examples of successfully developed and deployed systems, this report aims to initiate a discussion around the key ingredients for success that will enable AI/ML developments to have a measurable and positive presence at the bedside.

Introduction

This report provides an overview of the Health AI Deployment Symposium which took place in Toronto, Ontario on October 30th, 2019 as a joint collaboration between the Hospital for Sick Children (SickKids) and the Vector Institute (Vector). The Vector Institute is an independent, not-for-profit corporation dedicated to advancing artificial intelligence (AI) and excelling in machine and deep learning. Vector's vision is to drive excellence and leadership in Canada's knowledge, creation, and use of AI to foster economic growth and improve the lives of Canadians. SickKids Research Institute is one of the largest hospital-based research institutes in Canada, investing over $200 million annually to generate new scientific and clinical knowledge and working collaboratively to apply these new discoveries to improve the lives of children in Canada and around the world. By providing real-world examples of moving AI from the research lab to the clinic, this symposium provided Canadian healthcare leaders with the opportunity to learn from the "do's and don'ts" of integrating machine learning (ML) into healthcare from American partner institutions. Speakers from the University of Michigan, Kaiser Permanente, Duke University, University of Pennsylvania, and Johns Hopkins University provided a diverse array of success stories and associated challenges.

"Healthcare offers tremendous opportunities for AI to provide wide-ranging societal good and Ontario is uniquely positioned to take advantage," said Dr. Garth Gibson, CEO and president of the Vector Institute. Vector has more than 30 affiliated members engaged in healthcare-related projects. Vector has also accelerated efforts in the health-AI field over the past 12 months through focusing on three foundational interconnected health workstreams: World-Class Research, Widespread Application, and Analysis-Ready Accessible Data. Through the Widespread Application workstream, Vector is supporting projects that aim to deploy and integrate AI-enabled healthcare practices into hospitals across the province.[9] Known as "Pathfinder Projects", several were represented at our Deployment seminar via a moderated panel discussion.

The AI in Medicine Initiative (AIM) at SickKids is developing strategies for moving AI from *in-silico* to *in-patient* pediatric care. The initiative has a goal to deliver data-driven and personalized paediatric health-care for pediatric patients by providing a platform and tools to facilitate the implementation of AI into clinical practice. AIM's projects encompass machine learning, deep image recognition, natural language processing, virtual reality, robotics and more. The initiative has an integrated partnership of leading researchers, computer scientists and clinicians at SickKids and works closely with Vector.

Despite the world class research and the constant flood of innovative articles in the fields of AI and ML in healthcare (see Figure 1), only a small percentage of articles mention concepts related to implementation or deployment, let alone lead to actualized tools. If the current translational success rate remains unchanged, funding agencies and the public may begin to grow disappointed at the lack of measurable progress in actual translational success stories.

There are a variety of domain-specific reasons explaining why delivering AI products which impact clinical care has proven difficult in healthcare.[10] Issues regarding algorithm evaluation,[11] model maintenance, integration into EHR or hospitals' data warehouses,[12] as well as privacy and fairness considerations[13] have all led to project development times which are longer than in other fields.

Dr. Ronald Cohen, the CEO of SickKids, identified machine learning tools as equivalent to the iconic stethoscope in medicine in his opening remarks. The success of these tools will impact how physicians are trained in the future as well as the types of individuals who will want to pursue a career in medicine. He congratulated Dr. Anna Goldenberg, the Varma Family Chair in Biomedical Informatics and Artificial Intelligence at SickKids, for pushing clinical practice to establish the necessary feedback loop for project development between computational and healthcare experts. Dr. Goldenberg identified four areas where additional investment will help expedite project development: streamlined access to data, appropriate ethical guidelines, systems designed with human-computer interaction (HCI) in the loop, and a clear pathway to deployment.

In the remainder of this report we summarize the presentations of Jenna Wiens, Vincent Liu, Mark Sendak, Corey Chivers, Suchi Saria, and a group of presentations supported by Vector's Pathfinders program in the order in which they were presented at the symposium. A question and answer summary has been provided at the bottom of the speaker section where relevant. The symposium's agenda is available in the appendix.

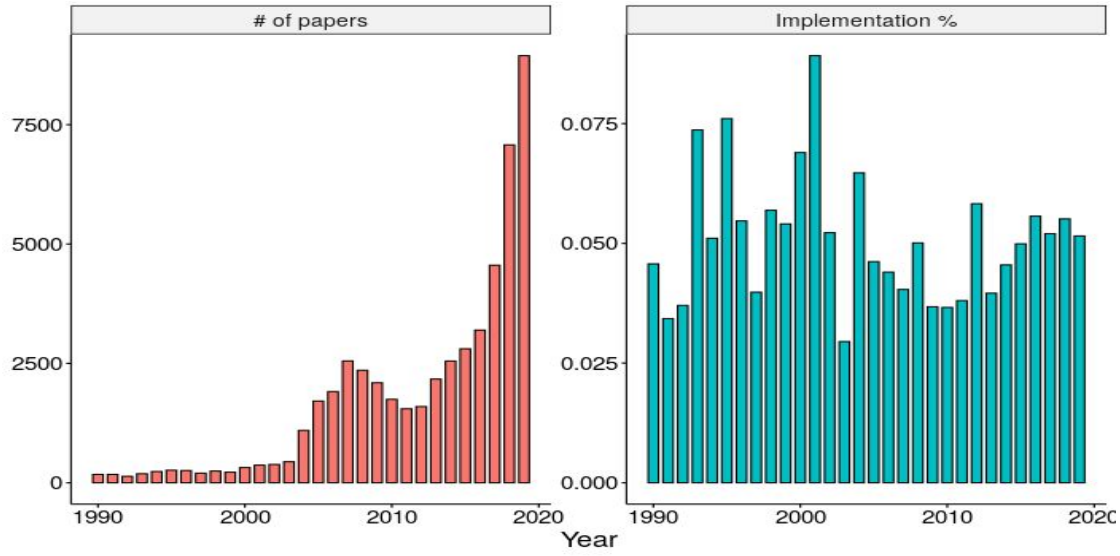

Figure 1: Number of the papers published in PubMed with a reference to "machine learning" or "artificial intelligence" anywhere in the article (a), and the percentage of which have the term "deployment" or "implementation" in them (b).

1. Accurately Predicting Healthcare-Associated Infections at Scale

Speaker: Jenna Wiens, University of Michigan
Discussant: Anna Goldenberg, Vector Institute, SickKids Research Institute

Dr. Jenna Wiens is a Morris Wellman Assistant Professor of Computer Science and Engineering (CSE) at the University of Michigan in Ann Arbor. She currently heads the Machine Learning for Data-Driven Decisions research group where she started translating ML into clinical practice. One of her group's primary research foci has been using ML tools to be able to detect hospital-acquired infections like *Clostridium difficile* Infection (CDI). These infections remain an ongoing clinical challenge to detect and cost the healthcare system billions of dollars leading to thousands of unnecessary deaths. By combining multiple types of data modalities, ML tools can provide clinicians with accurate pre-test probabilities of CDI thereby allowing clinicians to stratify patients into high-risk groups and expedite medical interventions[14].

Most algorithms used in the healthcare setting have an explicit (e.g., length of stay) or implicit (e.g., sepsis alert) time dimension. Dr. Wiens' work has shown that when predicting an implicit time-dependent outcome like hospital mortality, structuring the predictions with respect to an outcome-independent reference point is essential[15]. For example, predicting whether a patient will experience an event in the next 72 hours from the time of admission satisfies this requirement as the time of admission is an independent reference point to the outcome of interest. In contrast, making the same prediction 72 hours before the event has happened does not satisfy this requirement as a prediction can never be indexed to a future time point in a real-time setting. Furthermore, by using outcome-dependent timepoints the accuracy of the algorithm will appear inflated due to selection bias (e.g., the healthiest patients are discharged quickly).

Transfer learning represents one of the most important applied research successes in ML -- particularly in the fields of natural language processing and computer vision[16]. However transferring models from one hospital to another has been challenging due to a variety of strong hospital-centric features[17]. Dr. Wiens stressed the importance of leveraging hospital-specific features during model development and using *generalizable approaches* in contrast to *generalizable models* for multicentre studies. A "one-size-fits-all" model is unlikely to generalize across institutions due to challenges in mapping variables between hospitals, differences in EHR systems, hospital practices, and testing procedures. Hospital-specific CDI classifiers were therefore trained for the University of Michigan Hospital (UM) and the Massachusetts General Hospital (MGH)[18]. However pooling data across institutions can be useful for small hospitals where there is a paucity of data.

A consistent theme throughout multiple speakers presentations was that even though ML algorithms are often able to obtain good discriminatory ability, as measured by the AUROC, these algorithms often have a low Positive Predictive Value (PPV), which is the number of true positives divided by the number of predicted positives. In the case of the CDI tested at UM, the 0.82 AUROC translated to a 5.6% PPV at a fixed 95% sensitivity due to the relatively low prevalence of CDI[18]. Whether a tool in which approximately 17 in every 18 alerts is a false positive is valuable to the institution depends on the burden of alert fatigue, the necessity of having a high sensitivity, and the cost of intervention. In settings where the treatment is relatively cheap and noninvasive (e.g., probiotics) such a ratio may be acceptable. The current CDI algorithm tailored to UM has its own custom infrastructure with a web-service pull of EHR data occurring daily at 12am. The challenge of extracting data from Epic EHRs in real time was another theme many speakers raised. For example there is often a delay between the Epic Chronicle and Clarity databases followed by a further delay between Clarity and a hospital's research data warehouse.

The timeline for developing a CDI risk stratification tool took more than three years. First, training must occur on retrospective data over a long enough time period. Second, a silent period must be initiated to ensure that *in-silico* performance approximates real-time accuracy and algorithmic bias can be assessed (e.g. differential model performance between ethnicities). Third, a hospital-wide study will be carried out to assess the overall success of the risk tool. While lessons learned from past projects can speed up time to deployment and implementation, having a clinical champion at each institution was identified as one of the most important accelerators. An expedited process of one year for training and one year for validation was suggested as the lower bound for a project of this nature.

Questions (Anna Goldenberg and participants**) and Answers (**Jenna Wiens**)**

Q1. How and why did you decide to work on CDI?

Work started as a graduate student and was encouraged by clinicians and hospital leadership to continue to pursue this problem.

Q2. How does one assess what a sufficient level of performance is?

It will depend on the problem and the clinical use case/intervention. For example a 5% PPV will not work for assigning private rooms in a hospital, but may work for a condition with a non-invasive treatment.

Q3. What team do you have in place to carry out model development and assessment?

Students, clinical collaborators working on infectious disease problems, as well as experts in IT, health implementation and study design.

Q4. How is your tool integrated into Epic?

The tool does not currently push alerts through Epic. Instead, we envision it sending messages to a subset (e.g., infection prevention team) rather than all physicians.

Q5. What were the lessons learned from a multisite project?

A clinical champion at each institution is needed to be able to push forward progress and bring down barriers.

Q6. What is the minimum amount of data you need to train a model, and can transfer learning help?

It depends. In settings with seasonal trends, a minimum of one year of data (to capture seasonal effects) followed by one year of validation is advisable. Transfer learning may reduce data requirements for small hospitals.

Q7. Did you check for bias in your models?

We checked the model inputs for biases in race and sex.

2. Identifying Health-System Level Opportunities for Predictive Model Deployment

Speaker: Vincent Liu, Kaiser Permanente
Discussant: Muhammad Mamdani, St. Michael's Hospital

Dr. Vincent Liu is a Research Scientist with the Northern California Kaiser Permanente Division of Research with extensive expertise in medical informatics, data science, and acute severe illnesses like sepsis. He is also the Regional Director of Hospital Advanced Analytics leading a multidisciplinary team embedding real-time predictive models into clinical practice. Kaiser Permanente Northern California (KPNC) is an integrated healthcare delivery system with 21 hospitals and more than 200 medical clinics. This scale of healthcare provision allows KPNC to act as a "data science lab in the wild" with ML tools able to be rolled out across all hospital sites. Algorithmic evaluation is expedited through the organizations large number of stakeholders including more than 4 million healthcare members, 9 thousand physicians, and 22 thousand nurses. KP has been on the leading edge of medical informatics helping to establish computer databases for healthcare[19] as well as personalized medicine[20].

Bringing machine learning tools to life requires two components: data science and delivery science. While the ingredients of data science in healthcare are well understood (researchers, model builders, and IT experts), the delivery science piece receives less attention and is often more complicated requiring operational leadership and clinician support. Dr. Liu identified the life cycle of a predictive model of having five stages: prioritization, assessment, development, deployment, and evaluation -- although the cycle does not often proceed via a linear path.

The identification of high-value targets that have modifiable outcomes with hospital leadership helps to prioritize ML tasks. Assessing how well similar tools have performed in other hospitals, what the expected resources needed to bring a project to fruition would be, and whether it makes sense to build the tool from scratch or purchase it, can help identify the highest-value and feasible projects. While the pre-test probability for a successful medical intervention remains low for many medical devices and pharmaceuticals[21], it is unclear whether the same holds true for ML tools evaluated under an RCT framework. Future research will be critical to help determine whether algorithmic *in-silico* success translates into clinical impact across a wide variety of predictive models. Project development requires identifying the details of technical

implementation as well as how the model will be calibrated and internally validated. Alignment with the hospital's existing work-flow and the design of the end-user interface are critical questions to be addressed by the deployment team. Lastly, evaluation needs to be carried out to ensure that the expected outcomes are being achieved and that a system is in place for ongoing model maintenance.

After a tool has undergone statistical development targeting a specific prediction problem, health system deployment requires alignment across two fronts to ensure a seamless "technical-operational handoff". What platform is needed (technical) and where will it fit into workflows (operational)? What is model performance (technical) and how will end-users interact with risk scores (operational)? How will alerts be sent (technical) and how is reliability ensured (operational)? What is the performance during the silent period (technical) and is this model making a difference on the ground (operational)? The technical "go-live" necessarily occurs before the operational "go-live" as sufficient data needs to be collected before operational questions can be addressed. Having a world-class delivery science team is critical to addressing all of these challenges.

One of KPNC's earliest models was an early warning score to detect inpatient deterioration[22]. Their tool is completely integrated into Epic with all presentations and calculations of scores staying inside the system. The risk scores are calculated in a high-throughput manner and are first delivered to a virtual triage team before alerts are sent to frontline clinicians. These nurses apply human evaluation to the scores and talk with the rapid response teams who then initiate the local workflows. The rollout of this algorithm showed decreased inpatient mortality, greater "goals of care" discussion, and a reduction in risk-adjusted readmission rates (forthcoming publication).

The research team at KPNC has turned their attention to a variety of other healthcare problems amenable to ML support including diabetes mellitus treatment related hypoglycemia[23], mental health risks including suicide attempts and deaths[24], labour and delivery deterioration[25], incident HIV infection [26], and incident sepsis hospitalization[27]. Projects around mental health and HIV infections have additional considerations and sensitivities given the vulnerable populations they affect. The KPNC research team categorizes project development along a gradient ranging from an initial "scientific" exploration to an intermediate "technical" stage, and then lastly to the "operational" level. This provides a common linguistic framework to understand the development cycle. To help operational leadership make choices around the risk alert threshold, a trade-off between the detection ratio (inverse PPV) and sensitivity are provided showing the average number of alerts a tool will send each day to specific end-users.

Most data science teams try to develop an algorithm which targets the patient of interest (e.g. which patients have sepsis), but changing the target towards a resourcing or workflow issues can sometimes be more effective. If 100 patients need a service but there are only 20 units available to provide that service, then an effective algorithm could simply focus on finding the 20 patients who need the service the most. In contrast, if 100 patients are receiving a service but only 80 need it based on an algorithm, then the algorithm could reduce service delivery for the

potentially ineffective subgroup. For workflows that are highly variable, ML can be used to standardize aspects of treatment delivery. When it is appropriate to focus directly on the patient, the algorithm should provide personalized treatment recommendations at a specific point in their disease trajectory. The performance characteristics of the model needed for these three targets (resources, workflows, and patients) will differ. Ideal predictive models need to be contextualized in a clinical workflow to help augment or enhance a clinician's decisions rather than to only try to replace it.

Due to time constraints only a single question was asked to Dr. Liu around the parallels between the environments seen in lab services and those of healthcare analytics. Data science teams have a lot to learn from lab services especially around the establishment of rigorous measures to ensure reliability and standardized processes. As the number of models being used in a system grows, a holistic approach is needed to assess performance in the same way that multiple measurements of creatinine, lactate, and troponin are critical care clinical practice, but do not dictate clinical action on their own.

3. Integrating Deep Learning into Routine Clinical Care to Rapidly Detect and Treat Sepsis

Speaker: Mark Sendak, Duke Institute for Health Innovation
Discussant: Amelia Hoyt, The Hospital for Sick Children

Dr. Mark Sendak is the Population Health & Data Science Lead at the Duke Institute for Health Innovation (DIHI) where he manages an interdisciplinary team of data scientists, clinicians, and ML experts to solve clinical problems. DIHI uses a top-down/bottom-up approach for sourcing innovation and has a direct line of communication with senior health system leadership. Value creation at DIHI occurs across the spectrum of care including inpatient innovations, transition settings, and gaps-in-existing-care. The goal of machine learning tools is help shift the production possibility frontier to deliver more high quality and low cost services.

One of DIHI's most successful Early Detection and Deterioration projects is Sepsis Watch which was brought to fruition after 2.5 years and included a variety of stakeholders[28]. Sepsis Watch is also one of the few hospital-based ML systems that has been registered as a clinical trial[29]. One common challenge in any data-driven sepsis project is defining the label. In addition to the existing Systemic inflammatory response syndrome (SIRS) criteria, other characteristics including mortality and morbidity also helped to guide label assignment. Understanding where (inpatient vs emergency department) and when (development relative to admission time) sepsis occurs helped to guide model development[30]. While the PPV for the Sepsis Watch algorithm was impressive compared to similar tools, a custom workflow was nevertheless implemented as the Epic Best practices advisory notifications were largely being ignored by physicians. Instead a remote team was provided with the sepsis alerts and they decided whether or not to reach out

to the treatment team on the ground. This closely resembles the strategy used by KPNC's inpatient deterioration tool discussed above. However an IT solution for Sepsis Watch was built outside of Epic (similar to Dr. Wien's CDI tool) in order to have real-time data extraction and predictions as well as a custom user interface (UI).

The UI for Sepsis Watch has three stages: triaging, monitoring, and treatment. The remote assessment team is first alerted at the triage stage. Monitoring tracks updates about whether a nurse or physician has been contacted. Finally the treatment stage allows the application to track which tests and treatments have been administered to the patient. Tools like Sepsis Watch require ongoing maintenance and engagement even after deployment. Sending out weekly compliance reports as well as handling natural perturbations to data fields are required to ensure trust is established among clinicians and the model remains relevant and valid. Duke also has a reporting group that works on all quality-improvement projects. For example, doctors are provided with the successes and failures of the patients they took care of.

At the beginning of the symposium Dr. Goldenberg identified appropriate ethical guidelines in healthcare AI as an important framework which needs to be further developed. Dr. Sendak echoed this concern and highlighted four types of ethical challenges ML will surface in healthcare: patient awareness and consent, clinician awareness and consent, retrospective bias, and prospective bias. In order to address each of these issues, there should be a discussion amongst stakeholders and experts around establishing general frameworks. DIHI has a similar view to KP's lifecycle approach for model development by separating these projects into three phases. In Phase 1 there is a problem assessment where relevant workflows can be hypothesized. In Phase 2, the detailed design of both the model architecture and the workflow is developed. Lastly, in Phase 3 implementation and evaluation is carried out and future governance structures are put in place.

Various lessons from the business world can be used to help provide better language and methods for strategic development in our field. First, change management frameworks can be used to improve the organization as well as the technology[6]. By referring to these technologies as *augmented* rather than *artificial* intelligence, it helps to make clear that these tools are designed to assist clinicians. By describing algorithms as being "deployed" stakeholders may come to see these systems as autonomous and decontextualized whereas "integrating" tools puts the technology within existing norms and practices[7]. DIHI has existed since 2012 and in the first few years they were unable to develop and deploy ML projects in the 12-15 month timeframe that pilots were funded for. The primary roadblock that was overcome was having access to clean and high quality data in a timely fashion.

Questions (Amelia Hoyt and participants**) and Answers (**Mark Sendak**)**

Q1. For organizations at the start of this journey, what are your lessons learned for ensuring engagement is effective and sustained?

> First, frontline clinical engagement is needed to help to surface the problems that can be addressed by ML and to help champion progress. Second, operational engagement needs to have someone be responsible for curating and funding the project after it is

deployed. Third, IT needs to be involved from the start to determine what can be done within and outside of Epic as well as maintenance.

Q2. How have you trained your team members and the operational side to better understand machine learning?

Programs were built that directly taught medical students coding and strategies for evaluating technologies. For individuals with technical expertise it meant getting them talking to the clinical side.

Q3. How important is interpretability?

The dominant narrative in healthcare is that for high-stakes decisions interpretability is necessary but over the multiple years that Dr. Sendak's team has been operating there have been almost no bottom-up requests for model interpretability in the sense of variable importance. Rather, stakeholders have requested "context" when receiving model predictions. This is more broadly related to building trust, in which interpretability *per se* is not necessary[31]. For tools like Sepsis Watch, front-line users want to see vital signs and labs and how the patient has evolved over time.

Q4. What are the lessons learned when building your own data pipelines?

Some models can be implemented directly into Epic when internal data representations are sufficient. However for tools like Sepsis Watch custom infrastructure needed to be built. Ultimately it depends on how complicated the data extraction is and how often models need to be updated.

4. The Right Solution to the Wrong Problem? Lessons from Deploying an ML-based Early Warning System

Speaker: Corey Chivers, University of Pennsylvania
Discussant: Devin Singh, The Hospital for Sick Children

Dr. Corey Chivers is a Senior Data Scientist at Penn Medicine where he works with clinicians to take predictive healthcare solutions from the idea and experimentation phase to scaled production implementations. The data science branch at Penn Medicine was established in 2014 and one of its earliest projects was an ML-based early warning system (EWS) for sepsis to reduce mortality. This system was one of the largest ML-based EWS which was deployed and prospectively evaluated across two large academic hospitals [32–34]. Prior to full-blown sepsis presentation or a positive blood culture drawn (ie. Lactate > 2.2 or BPS <90), there exists a clinical state that Dr. Chivers refers as preclinical detectability which patients transition into from an original stable state. A supervised learning model was trained to detect patients who would go on to develop sepsis in this preclinical detectability state, with outright sepsis cases excluded. The ML model for the EWS was a decision tree based algorithm which was trained on

various features extracted from observations (eg. labs and test results) within a 12 hour window prior to sepsis onset.

As a part of the model development and also in an attempt to make the model re-usable, they created a data science pipeline, called ***Penn Signals***, which infuses and integrates multi-modal EHR data from different sources and converts them into time-series signals. The data science team could therefore get access to both batch and realtime data for their ML model prototyping and deployment. After a design period and technical troubleshooting achieved reasonable results (PPV of almost 30%), the team proceeded to two phases of validations: 1) a 6 month silent period and 2) a 6 month alert period. The only difference between the two phases was that no alert was sent for a silent period. The alerts were sent to the care team, nursing coordinator, and providers in the form of a secure text message. They were also provided with a context plot of the vital signs. To further investigate the clinical utility of the model, the team compared changes in clinical outcomes between silent and alert phases and, unfortunately, there was no statistically significant difference except for the time to ICU admission.

This result was all the more surprising because the model performance seemed to be high during the silent trial. The team therefore surveyed clinicians and nurses to assess their perceptions of the utility and impact of the EWS 2.0 on patient care. The survey results indicated that users, in general, were not supportive of the model or impressed with its benefits. There are three reasons that Corey and his team believe the system was not improving patients care. First, the dataset was imbalanced as the number of patients who developed sepsis following admission but were not diagnosed at the time of admission was small. Second, the majority of patients that were becoming septic at the time of the alerts were already suspected of being so by the care team. Lastly, there was no intervention that was consistently carried out for patients at the 'pre-sepsis' state when alerts were sent.

The evidence indicated that choosing the set of right problems to address is an important piece in moving AI for health from research to deployment. Dr. Chivers and his colleagues, therefore, created a very short questionnaire called the "Predictive Healthcare Madlib" that help them remain focused on right predictive solutions for the right problem that can result in an improved care plan:
As a [decision maker], If I knew [information], I would do [intervention], to improve [measurable outcome]

Two ongoing projects were launched as a result of the outcome of the Predictive Healthcare Madlib: (1) to detect individuals who may benefit from palliative care, and (2) to identify patients ready for transition from ventilator to spontaneous breathing. Both projects are in the preliminary stage and their successful integration into the hospital setting would result in a clinical care improvement. Dr. Chivers also raised the issue of how to formulate the model's loss function to account for the cost of interventions and errors in event detection. This information can in turn be used to determine what are the satisficing metrics needed for model usefulness.

It is important for the community to understand the problems they are trying to solve and frameworks such as the Healthcare Madlib can help clarify roles, actions, and expected outcomes. It is also useful to leverage decision theory rules prior to building the model in order to understand the ML performance that would be required in order to expect positive outcomes over alternatives (i.e., status quo, treat all, treat none). The talk reinforced the view that the successful integration of ML models into healthcare is as, if not more, challenging than the task of developing these models.

Questions (Devin Singh and participants**) and Answers (Corey Chivers)**

Q1. Is there any particular lesson that you learned from your past experiences that you would like to share with people who are at an early stage of ML model development in health with respect to working with retrospective data? And what would you do differently if you go back to the beginning of the project?

> Real-time data is very different from research data stored in the hospital data warehouses. Getting data and making a decision in real-time requires incorporating additional latency for feature extraction and inference as well as online noise filtering and unit standardization. Unlike offline data, real-time data comes with different timestamps. These are the essential factors that should be taken into account by the project team right at the beginning and prior to any ML model development.

Q2. How did the team choose to use text message nudges as opposed to a phone call or any other alert within Epic Best Practice alerts?

> Given the short timeline of the project, text messages were the quickest and most feasible way to deliver outcomes and alerts to clinicians' hands.

Q4. How do differential costs between false positives and negatives affect the applicability of the decision theory process the team is using in order to have a more structured decision?

> Based on decision theory framework, we are looking at a population level where all individuals are grouped and treated together. Although it seems a reasonable approach to pursue, it might fail in capturing variation among the population. But it's important to note that it is a joint human-machine collaboration and the final decision will be made by clinicians. So, hopefully, they would be able to handle those diverse set of cases in the outcome on the confusion matrix. However, for a fully automated decision making which results in high risk intervention, there should be additional considerations in place around the distribution of the outcome.

Q5. Could you explain more about the methodology your team is approaching in sampling and modeling of the cost-benefit trade-offs and which expertise are required?

> There are already well known methods with respect to the implementation of expected utility maximization. What is less established is how to define and value different terms (e.g., cost of interventions). In our project, we always use a cost function to formalize the objectives and a stakeholder-derived framework. From this, we can start to ask "In what range would the unknown quantities need to be in order for us to choose the model of some alternative for decision making?"

Q6. With regards to the outcome of the model, did you look into clinicians' experiences in using the tool?

There was only a provider survey that was primarily designed to ask clinicians for their perception of the tool. However, it is a great idea to have a tool or survey that asks about the usefulness of the alert as well as user experience.

5. Takeaways from 9 years of collaborations across more than 10 service lines

Speaker: Suchi Saria, Johns Hopkins University
Discussant: Mjaye Mazwi, The Hospital for Sick Children

Dr. Suchi Saria is the John C. Malone assistant professor of computer science at the Whiting School of Engineering, health system informatics at the Johns Hopkins University School of Medicine, and health policy and management at the Bloomberg School of Public Health. Her research work in healthcare AI spans a wide variety of projects, and she has done important work to demonstrate how diverse signals can be integrated in ML systems to make early detection for sepsis possible [35].

Dr. Saria illustrated the passage from science to delivery as a tunnel made up of three main zones: *development*, *launching* and *business*. The *development* zone is where most researchers and scientists spend the majority of their energy, time, and funds. Ideation, model development and evaluation on retrospective data, along with scientific publications are all conducted in this zone. Researchers who have positive findings and media attention will begin to work in the *development* zone as part of career development. As you move down the tunnel and get closer to its neck, you will reach the *launching* zone where the ML model is transferred from the virtual to the real such as designing the workflow, educating the users, estimating the infrastructure, and integration into a real healthcare setting. Dr. Saria calls this stage a “dead zone” as the researchers and developers might give up if they see the performance of the model decreasing in a natural setting with real-time data. However, if you manage to pass by this zone the tunnel will widen out and you can reach the survival zone which is the *business* zone. It is a less interesting zone for researchers and scientists but could end up being a great success in terms of the return on both private and public investment. This zone requires reporting results, continual change management, monitoring, maintenance, and model improvement. Dr. Saria has been working on the computational-detection of sepsis for almost 7 years now. One of her models was deployed in the background in early 2017 and integrated in an Epic workflow. Her team spent almost 18 month to get to the point where clinicians were fully satisfied with the tool. Dr. Saria’s team for productionizing the sepsis prediction ML model included herself and nine other individuals with various skill sets ranging from data science to devOps and user design engineers.

There are different factors that should be taken into consideration in moving from development to integration of ML models in health settings. These factors can be classified into two main categories: clinical and workflow. Clinical considerations that need to be taken into account

include the frequency of the underlying event, the importance of the precision for the prediction problem, the harms of missing a patient, and the burden of over-treatment. Workflow considerations are made up identifying the user (patients vs. care providers), adoption strategies, and hedging financial risks. In addition to the above considerations, teams should also ensure the accuracy of their models over time and across settings.

Dr. Saria also raised the concern regarding the robustness and transformability of ML models against changes in setting and time once they are deployed to production. The example of adversarial inputs was given as a warning to how fragile some ML systems can be[36]. Another example provided was a deep learning model for detecting pneumonia in chest radiographs where in 3 out of 5 natural comparisons, performance on chest X-rays from outside hospitals was significantly lower than the original hospital system[37]. In a similar study, Dr. Saria's team also observed performance degradation for a model trained on data from 2011-2013 and tested on 2014[38]. Dataset and/or covariate shifts are common problems in predictive modeling that occur in most practical applications.[38–40] Training data that have structural biases will reduce the effectiveness of predictive models and pose an ethical challenge, especially healthcare, where decisions can be life or death ones. Shifts in the data such as provider practices should therefore be corrected where possible.[38,40–43]

As ML projects move from small scale research projects to large scale clinical deployment, a large amount of infrastructure is required to support its production. Additional steps to avoid model staleness are required with a feedback loop between monitoring systems and domain experts in place. The actual infrastructure requirements needed to scale an operation remains an ongoing area of work and research.

Questions (Mjaye Mazwi and participants) and **Answers** (Suchi Saria)

Q1. How do you imagine that the incentive structure could be changed in order to get the effective model to the point of care given that it is an ecosystem level intervention in the provision of healthcare? Applying ML in healthcare requires the ability to modify a workflow, understand use cases, adapt to changing conditions, and develop custom infrastructures.

> In order to produce an effective model, there should be a team of people with diverse sets of knowledge and expertise including human factor engineers, implementation scientists, data engineers, people familiar with clinical transformation, and EHR workflow. Building and sustaining an effective model requires a central institutional mandate. Projects which focus solely on delivering "innovation" will not be able to sustain themselves past launching the project. Successful AI products in the market have in place significant resources to both productionizing and prototyping. In academia, this could be achieved through collaboration with industry partners.

Q2. In your 9 years of experience across different institutions, what is the biggest barrier?

> The biggest barrier is talent in terms of hiring, maintaining, and sustaining it. A good team knows how to work together and understands all of the pieces in the pipeline. Dr. Liu agreed with Dr. Saria's in regards to having a central mandate as resources need to be pulled across institutions to run a model. To have success with multiple models, a

team needs to be able to speak the same language and understand how an iterative development process works.

Q3. What is the critical success factor for a new organization needs before embarking on a data science journey?

Having a task force, governance council, and infrastructure team. As well as being aware of the challenges in the different phases a project takes on. Implementation and deployment takes a longer time than prototyping and requires patience and maintenance.

6. Pathfinder Projects

Alison Paprica, VP of health strategy partnership at Vector institute, introduced Pathfinder Projects as small-scale efforts designed to take research that is nearly ready to go into clinical practice. With technical and resource support from the Vector Institute, they each bring together a multidisciplinary research team to tackle an important healthcare problem or opportunity using ML and AI more broadly. Each project was chosen for its potential to help identify a “path” through which world-class ML research can be translated into widespread benefits for patients.

Medly Project - Application to Remotely Monitor patients with congestive Heart Failure

Dr. Heather Ross, a cardiologist at the Peter Munk Cardiac Centre (PMCC), Professor of Medicine at the University of Toronto, and Director of the Cardiac Transplant Program at Toronto General Hospital opened her talk with a statistic on Heart Failure (HF) in Canada. About one million Canadians (1 in 5) over the age of 40 are diagnosed with HF and half of Canadians have either experienced heart failure themselves or as a care-giver. The overall cost in terms of days in hospital is tremendous (1.4 million days with the average patient spending 26 days in hospital) as well as the financial cost to the Canadian healthcare system ($2.3 billion). Ensuring optimal care to this growing population of patients is one of the most vexing challenges facing healthcare professionals working in cardiovascular medicine. In Canada, HF patients typically see their cardiologists twice per year. Between these visits, doctors remain unaware whether symptoms are worsening and patients are left guessing about the health of their heart. Medly is a digital tool that easily allows patients to measure their weight, blood pressure and heart rate and sends this information electronically to a nurse who carefully monitors their care. It seeks to cover patients in between their visits and provide them with self-care and coordinated clinical support, all without leaving their homes. Medly runs this data through a clinically-validated algorithm which provides instant feedback to patients and their clinicians on a daily basis. The Medly system is part of the PMCC's digital cardiovascular health platform (DCHP) and has been integrated into the centre's care for heart failure patients. The Medly app was co-developed by Dr. Heather Ross (PMCC) and Dr. Joseph Cafazzo’s team, a biomedical engineer, researcher, and educator at eHealth Innovation, a partner of University

Health Network. One of the issues with the current Medly app is that there are many false positives. The team has now established an extension to the Medly app, called Medly AI, which will be used to explore how ML can help to reduce the frequency of false positives without sacrificing patient safety. Medly AI will use different data features (from genomics to EHR) from DCHP along with daily self-measurements which feed into an AI engine.

Dr. Bo Wang, the Lead Scientist of the Artificial Intelligence Team for Peter Munk Cardiac Centre at University Health Network and also a Faculty member at Vector Institute explained briefly about the Machine Learning component of the project. He pointed out two approaches that will be used in Medly AI as follows: 1) a decision tree binary classifier (Random Forest), and 2) recurrent neural network deep learning model.

Rule-based Medly currently has a Health Canada (HC) class 2 status under the Medical Device Quality Management System. And unlike other telemonitoring devices, where one nurse is assigned to 40-60 patients, Medly enables that single nurse to manage more than 350 patients using the Medly app. Medly is also part of a QI project to reduce HF-caused hospitalization at Sunnybrook, where it has been shown to reduce hospitalizations. Dr. Ross added that the number of patients with HF is expected to increase by about 25 percent over the next 20 years. Hence, novel mechanisms are needed to treat patients. Using mHealth to treat HF will enlist the help of the largest healthcare workforce.

Tick Identification to combat Lyme Disease

Vanessa Allan, Chief Medical Microbiology at Public Health Ontario (PHO) explained about the rapid increase of Lyme disease in Ontario and Canada due to tick bites. Blacklegged ticks are the only ticks in Ontario known to carry *B. burgdorferi*, the bacteria that causes Lyme disease. While not all blacklegged ticks carry *B. burgdorferi*, a bite from one is of more concern than a bite from other tick species which do not carry the bacteria. Lyme disease can be prevented if antibiotics are given within 72 hours of a tick bite in an area of disease activity. Because of the lack of expertise in tick identification in Ontario, the ability to detect the tick type within 72 hours has been limited. The current process is that an individual removes the tick from their body, puts it in a plastic bag, goes to a clinician, who sends it to a PHO lab, and then finds out the results a month later. Ideally the tick should be identified by a clinician, and the lab submission is primarily designed for surveillance purposes. However, clinicians do not have required skills to detect ticks. Part of Dr. Allen's project aim is to build an AI model that will empower the public to take care of their health within a window where they can get effective treatment. There is an interdisciplinary team of physicians, microbiologists, laboratory technologists, engineers, scientists, entomologists, and public health specialists involved. The team has currently developed a computer vision model to differentiate between the two common tick species normally found in Ontario. The model has been trained on an image bank that has 13 thousand tick images taken from PHO labs. The test set accuracy of the model is greater than 90%. As a first deliverable, the team has already developed a web app that professionals at PHO will use to identify whether or not a tick is a blacklegged tick. The long-term goal is to create an app that

anyone can use to simply take a photo of a tick. Once the app identifies the species, it will provide advice. In addition to empowering the public in management following tick bite, the app will improve surveillance of ticks and their geographic distribution which results in laboratory efficiency.

Early Warning System for General Internal Medicine

Dr. Amol Verma, internist physician and clinical scientist mentioned that at St. Michael's Hospital, a 450 bed health centre, an average of 8% of General Internal Medicine (GIM) in-patients will die or be transferred to Intensive Care (ICU) at the hospital. Unrecognized deterioration is the most common root cause of unplanned ICU transfer. An EWS may therefore help for early detection of deterioration and interventions which is currently deployed in the majority of hospitals in the UK (>75%). The challenge with existing EWS tolls is that they recognize but to do predict deterioration and suffer from high false alarm rates. The team set out to build CHARTwatch, an EWS based on AI, to reduce mortality and improve the quality of care in GIM inpatients. The model was built using approximately 10 years of historical data including more than 800 data elements from the hospital's EHR. The dataset included 20 thousand historical encounters where each encounter was segmented into 6-hour time intervals, and the model was built to make a prediction at the end of each time interval. An out-of-time validation approach was used to evaluate the performance of the model where the training, validation, and test sets included encounters from 2011 to 2017, 2018, and 2019, respectively. As mentioned above, the performance of CHARTWatch was tested out-of-time on 4-months of held-out data from May-Aug 2019 and compared with NEWS and clinicians. The model was compared to almost 4K daily realtime predictions made by patient nurses and attending physicians. All clinicians' predictions were less sensitive than either NEWS or CHARTWatch. NEWS has the lowest PPV while MDs have slightly higher PPV than CHARTWatch. However CHARTWatch had the highest AUROC. The time interval between the alarm and actual event is about 1.5 days, which provides a good amount of advanced notice to make interventions. This is a multi-stakeholder project including a team of local full-time data scientists, researchers based at the University of Toronto and Vector Institute, clinical implementation committee, evaluation committee, patients' advisors and out of that group, there is an implementation team as they are moving more toward deployment.

This project is a huge team effort and the team built a care pathway in order to improve the quality of care for patients through reducing non-palliative death in hospital as well as reducing time-to-palliative and -intensive care consults. Dr. Varma pointed out that the model predictions will go to physicians and nurses from GIM as well as the palliative care team. Right upon successful development of the model with promising results, the team is in transition to a silent period phase where the plan is to revisit with physicians and inquire what they would do differently or how they would intervene if they were given a prompt prediction on high risk patients in a hope that it would improve current care pathway. Following a silent period phase, the team was interested in running an RCT but this would require approximately 25K patients,

which would require almost 5 years. The team is therefore considering two alternative approaches such as running a cohort study and/or time-series analysis to evaluate the EWS impact.

Coral Review Project - Computer vision System to support Diagnostic Imaging

Dr. Antonio Sze-To is a postdoctoral fellow at the University of Waterloo and a member of Knowledge Inference in Medical Image Analysis (Kimia Lab) led by Dr. H.R. Tizhoosh, a Faculty Affiliate at the Vector Institute.

In this talk Dr. Sze-To explained how their AI tool, Insignio, detects pneumothorax in X-Ray images. Their model was developed with the collaboration of University Health Network (UHN) and will be integrated within Coral Review. Coral Review, a software solution developed at UHN, is a peer learning tool used by clinicians in diagnostic imaging to support the continuous quality improvement of radiology practice. Coral Review has been implemented at a number of hospitals across Ontario and enables an anonymous peer review of a medical imaging diagnosis, as well as image quality.

Pneumothorax is a life-threatening emergency that is very likely to be detected from X-ray images. Training radiologists is expensive and time-consuming with current imaging volumes leading to delays in X-ray reviews. An automated method of prioritizing X-rays with pneumothorax may reduce time to treatment. To bring more regularity and efficiency into the system, Antonio brought up an image search solution that scans through thousands of existing medical images (i.e., X-rays) for ones similar to a patient's and recommend a diagnosis to the attending physician through majority voting. Insignio is composed of two parts: a deep convolutional neural network that learns features of images, and an image search where the model retrieves the most similar training images to a given image. So far the team has evaluated the model on publicly available chest X-ray images [44–46] through K-fold cross validation with promising results (AUROC > 74%). After research ethics board approval, the team will integrate Insignio within Coral review to find similar looking images from past cases and offer suggested diagnoses -- while still leaving the final decision to doctors. And as a future direction for this study, Insignio can be easily adjusted for other chest diagnostics in addition to Pneumothorax.

Artemis Project - Predictive Analysis for Newborns in ICU

Dr. Carolyn McGregor AM is the Research Excellence Chair in Health Informatics and a two-time Canada Research Chair based at Ontario Tech University, Canada. The Artemis Project is a predictive analytics platform that applies ML to help physicians with the critical care of newborns. Dr. McGregor started her talk with a real example where a preterm infant in Australia was suspected with sepsis on day 10 of life and antibiotics were recommended.

However, unfortunately, 14 hours later, the baby died. Dr. McGregor's team observed a similar scenario in a hospital in Ontario where the baby survived and they were able to capture all the bedside signals. Variations in indicators like heart rate or breathing are signs of infections in infants. Should such signs occur, Artemis will alert physicians who will interpret the data and decide next steps. Artemis is being developed in partnership with McMaster Children's Hospital and Southlake Regional Health Centre. Once fully implemented, the Artemis system will monitor infants in neonatal intensive care units (NICU)s, alerting clinicians when sepsis develops before it would otherwise be clinically apparent. Ultimately, Artemis aims to reduce mortality, morbidity and average length of stay in NICUs.

Artemis was trained on retrospective heart rate and respiration data collected at a 30-second sampling rate from 1150 infants. The initial Artemis deployment was for 500 infants for real time data capture and prediction. Next, Artemis was deployed to capture data over 2 years from 80 bedspaces in the USA for 250 infants. The system is currently being deployed in two hospitals in Ontario (Southlake Regional Health Centre in Newmarket and the NICU at McMaster Children's Hospital in Hamilton) for real-time data streaming from over 400 infants. Artemis is running continuously in the background and makes predictions in the silent period. The team has deployed a cloud infrastructure in the Centre for Advanced Computing at Queen's University for storing and analyzing data in realtime. The infrastructure is built to work in real-time at scale.

At the end of her talk, Dr. McGregor shared the lessons she and her team learned during deployment. She insisted on the fact that deployment is not just about the model and there are many other elements that are playing an important role. Factors such as beginning with the end goal in mind, real-time data collection from remote bedside monitors, feature extraction, Al and analytics, as well as how and when to communicate with clinicians in terms of predictions should be considered. Once Artemis is fully implemented, Dr. McGregor is planning for the tool to be expanded beyond sepsis detection. The team is also deploying pilot studies into hospitals in India, Australia and South Africa.

References

1. Rajpurkar, P. *et al.* CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning. *arXiv [cs.CV]* (2017).
2. Gulshan, V. *et al.* Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. *JAMA* **316**, 2402–2410 (2016).
3. Shaw, J., Rudzicz, F., Jamieson, T. & Goldfarb, A. Artificial Intelligence and the Implementation Challenge. *J. Med. Internet Res.* **21**, e13659 (2019).
4. Bauer, M. S., Damschroder, L., Hagedorn, H., Smith, J. & Kilbourne, A. M. An introduction to implementation science for the non-specialist. *BMC Psychol* **3**, 32 (2015).
5. Greenhalgh, T. *et al.* Beyond Adoption: A New Framework for Theorizing and Evaluating

Nonadoption, Abandonment, and Challenges to the Scale-Up, Spread, and Sustainability of Health and Care Technologies. *J. Med. Internet Res.* **19**, e367 (2017).
6. Kotter, J. P. *Leading Change*. (Harvard Business Press, 2012).
7. Mateescu, A. & Elish, M. C. AI in Context: The Labor of Integrating New Technologies. *Data&Society report, available at https://datasociety. net/wpcontent/uploads/2019/01/DataandSociety_AIinContext. pdf, last time accessed by authors* (2019).
8. Wiens, J. *et al.* Do no harm: a roadmap for responsible machine learning for health care. *Nat. Med.* (2019) doi:10.1038/s41591-019-0548-6.
9. Vector Institute kicks off series of Pathfinder Projects focused on health AI adoption. *Vector Institute for Artificial Intelligence* https://vectorinstitute.ai/2019/05/06/vector-institute-kicks-off-series-of-pathfinder-projects-focused-on-health-ai-adoption/ (2019).
10. Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G. & King, D. Key challenges for delivering clinical impact with artificial intelligence. *BMC Med.* **17**, 195 (2019).
11. Craig, P. *et al.* Developing and evaluating complex interventions: the new Medical Research Council guidance. *BMJ* **337**, a1655 (2008).
12. Hersh, W. R. *et al.* Caveats for the use of operational electronic health record data in comparative effectiveness research. *Med. Care* **51**, S30–7 (2013).
13. Chen, I., Johansson, F. D. & Sontag, D. Why Is My Classifier Discriminatory? in *Advances in Neural Information Processing Systems 31* (eds. Bengio, S. et al.) 3539–3550 (Curran Associates, Inc., 2018).
14. Wiens, J., Campbell, W. N., Franklin, E. S., Guttag, J. V. & Horvitz, E. Learning Data-Driven Patient Risk Stratification Models for Clostridium difficile. *Open Forum Infectious Diseases* vol. 1 (2014).
15. Sherman, E., Gurm, H., Balis, U., Owens, S. & Wiens, J. Leveraging Clinical Time-Series Data for Prediction: A Cautionary Tale. *AMIA Annu. Symp. Proc.* **2017**, 1571–1580 (2017).
16. Pan, S. J. & Yang, Q. A Survey on Transfer Learning. *IEEE Trans. Knowl. Data Eng.* **22**, 1345–1359 (2010).
17. Wiens, J., Guttag, J. & Horvitz, E. A study in transfer learning: leveraging data from multiple hospitals to enhance hospital-specific predictions. *J. Am. Med. Inform. Assoc.* **21**, 699–706 (2014).
18. Oh, J. *et al.* A Generalizable, Data-Driven Approach to Predict Daily Risk of Clostridium difficile Infection at Two Large Academic Health Centers. *Infect. Control Hosp. Epidemiol.* **39**, 425–433 (2018).
19. Collen, M. F. & Davis, L. S. Computerized medical records in multiphasic testing. *Computer* vol. 6 23–28 (1973).
20. Garfield, S. R. The delivery of medical care. *Sci. Am.* **222**, 15–23 (1970).
21. Prasad, V. & Cifu, A. Medical reversal: why we must raise the bar before adopting new technologies. *Yale J. Biol. Med.* **84**, 471–478 (2011).
22. Escobar, G. J. *et al.* Piloting electronic medical record--based early detection of inpatient deterioration in community hospitals. *J. Hosp. Med.* **11**, S18–S24 (2016).
23. Karter, A. J. *et al.* Development and Validation of a Tool to Identify Patients With Type 2 Diabetes at High Risk of Hypoglycemia-Related Emergency Department or Hospital Use. *JAMA Internal Medicine* vol. 177 1461 (2017).
24. Simon, G. E. *et al.* Predicting Suicide Attempts and Suicide Deaths Following Outpatient Visits Using Electronic Health Records. *Am. J. Psychiatry* **175**, 951–960 (2018).
25. Escobar, G. J. *et al.* Automated early detection of obstetric complications: theoretic and

methodologic considerations. *Am. J. Obstet. Gynecol.* **220**, 297–307 (2019).
26. Marcus, J. L. *et al.* Use of electronic health record data and machine learning to identify candidates for HIV pre-exposure prophylaxis: a modelling study. *Lancet HIV* **6**, e688–e695 (2019).
27. Liu, V. X., Escobar, G. J., Chaudhary, R. & Prescott, H. C. Healthcare Utilization and Infection in the Week Prior to Sepsis Hospitalization. *Crit. Care Med.* **46**, 513–516 (2018).
28. Sendak, M. P. *et al.* Sepsis Watch: A Real-World Integration of Deep Learning into Routine Clinical Care (Preprint). doi:10.2196/preprints.15182.
29. Implementation and Evaluations of Sepsis Watch - Full Text View - ClinicalTrials.gov. https://clinicaltrials.gov/ct2/show/NCT03655626.
30. Futoma, J. *et al.* An Improved Multi-Output Gaussian Process RNN with Real-Time Validation for Early Sepsis Detection. *arXiv [stat.ML]* (2017).
31. Sendak, M. *et al.* 'The Human Body is a Black Box': Supporting Clinical Decision-Making with Deep Learning. *arXiv [cs.CY]* (2019).
32. Ginestra, J. C. *et al.* Clinician Perception of a Machine Learning-Based Early Warning System Designed to Predict Severe Sepsis and Septic Shock. *Crit. Care Med.* **47**, 1477–1484 (2019).
33. Giannini, H. M. *et al.* A Machine Learning Algorithm to Predict Severe Sepsis and Septic Shock: Development, Implementation, and Impact on Clinical Practice. *Crit. Care Med.* **47**, 1485–1492 (2019).
34. Blum, J. M. Beware of the Magic Eight Ball in Medicine. *Crit. Care Med.* **47**, 1650–1651 (2019).
35. Henry, K. E., Hager, D. N., Pronovost, P. J. & Saria, S. A targeted real-time early warning score (TREWScore) for septic shock. *Sci. Transl. Med.* **7**, 299ra122 (2015).
36. Goodfellow, I. J., Shlens, J. & Szegedy, C. Explaining and Harnessing Adversarial Examples. *arXiv [stat.ML]* (2014).
37. Zech, J. R. *et al.* Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. *PLoS Med.* **15**, e1002683 (2018).
38. Schulam, P. & Saria, S. Reliable decision support using counterfactual models. *Adv. Neural Inf. Process. Syst.* (2017).
39. Lum, K. & Isaac, W. To predict and serve? *Significance* **13**, 14–19 (2016).
40. Subbaswamy, A. & Saria, S. Counterfactual Normalization: Proactively Addressing Dataset Shift Using Causal Mechanisms. in *UAI* 947–957 (2018).
41. Subbaswamy, A., Schulam, P. & Saria, S. Preventing Failures Due to Dataset Shift: Learning Predictive Models That Transport. in *Proceedings of Machine Learning Research* (eds. Chaudhuri, K. & Sugiyama, M.) vol. 89 3118–3127 (PMLR, 2019).
42. Schulam, P. & Saria, S. Can You Trust This Prediction? Auditing Pointwise Reliability After Learning. in *Proceedings of Machine Learning Research* (eds. Chaudhuri, K. & Sugiyama, M.) vol. 89 1022–1031 (PMLR, 2019).
43. Saria, S. & Subbaswamy, A. Tutorial: Safe and Reliable Machine Learning. *arXiv [cs.LG]* (2019).
44. Irvin, J. *et al.* Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. *arXiv preprint arXiv:1901. 07031* (2019).
45. Johnson, A. E. W. *et al.* MIMIC-III, a freely accessible critical care database. *Sci Data* **3**, 160035 (2016).
46. Wang, X. *et al.* ChestX-Ray8: Hospital-Scale Chest X-Ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2017)

doi:10.1109/cvpr.2017.369.
47. Brajer, N. *et al.* Prospective and External Evaluation of a Machine Learning Model to Predict In-Hospital Mortality of Adults at Time of Admission. *JAMA Network Open* vol. 3 e1920733 (2020).
48. Sendak, M. P. *et al.* Sepsis Watch: A Real-World Integration of Deep Learning into Routine Clinical Care. *JMIR Preprints* **15182**, (2019).
49. Kang, M. A. *et al.* Real-Time Risk Prediction on the Wards: A Feasibility Study. *Crit. Care Med.* **44**, 1468–1473 (2016).

Appendix

A1. Health AI Deployment Symposium Agenda Agenda

8:30 AM - Breakfast and Registration

9:30 AM - Welcome and Introduction

Garth Gibson, President and CEO, Vector Institute

Ronald Cohn, President and CEO, The Hospital for Sick Children

Anna Goldenberg, Associate Research Director, Health, Vector Institute and Senior Scientist, SickKids Research Institute

9:50 AM - Accurately Predicting Healthcare-Associated Infections at Scale

Dr. Jenna Wiens, University of Michigan

10:25 AM - Identifying Health-system Level Opportunities for Predictive Model Deployment

Dr. Vincent Liu, Kaiser Permanente

11:00 AM - Coffee Break

11:30 AM - Integrating Deep Learning into Routine Clinical Care to Rapidly Detect and Treat Sepsis

Dr. Mark Sendak, Duke University

12:05 PM - Lunch & Networking

1:05 PM - The Right Solution to the Wrong Problem? Lessons from Deploying an ML-based Early Warning System

Dr. Corey Chivers, University of Pennsylvania

1:40 PM - Talk 5

Dr. Suchi Saria, Johns Hopkins University

2:15 PM - Coffee Break

2:45 PM - Rapid Fire Pathfinder Project Presentations

Early Warning System for General Internal Medicine (St. Michael's Hospital)

Medly (University Health Network)

Coral Review Project (University of Waterloo, University Health Network)

Tick Identification to Combat Lyme Disease (Public Health Ontario)

Artemis Project (Ontario Tech University)

3:45 PM - Closing Remarks

A2. Better understanding a model's "silent" period and mode

Many speakers stressed the need to ensure that the computational validation of models aligns with how data will be received in a real-time environment. To use the terminology of this report, it is important that *in-silico* validation approximates *in-patient* care. However no amount of carefully designed backtesting on retrospective data will ever be able to fully confirm that a model will be able to function in real-time in a clinical care setting. It is for this reason that all of the presented projects at the symposium went through a "silent period" where the ML algorithm was set to "silent mode". Setting a model to silent mode means that its output does not directly impact clinical care. The silent period is the span of time when a model in silent mode is evaluated. There is variation in whether clinicians are involved in evaluating the silent period, with some having no interaction (*in-hospital mortality*),[47] while in others there can be clinician feedback (*Sepsis Watch*).[48] Other examples of silent periods to evaluate performance include the e-CART model to detect cardiac arrests and the EWS 2.0 system to detect sepsis.[33,49]