Mind the Gap: Enabling AI Deployment in Health

November 22, 2022

In November 2022, the Vector Institute held a symposium entitled, **Mind the Gap - Enabling AI Deployment in Health**. The symposium brought together leading experts in the field of AI and healthcare to share insights and lessons learned from real world examples of AI implementation in health settings. Examples showcased encompassed projects that have been deployed in health institutions, including Vector Pathfinder Projects which were led by various Ontario health institutions and supported by Vector.

Vector President & CEO:

“The world needs these innovations.”

Dr. Garth Gibson, Vector’s President and CEO, opened the symposium by highlighting Vector’s work “with more than 30 hospitals, over 30 other health and life sciences organizations, and more than 20 universities across the province.” Gibson recognized the federal government’s $443 million commitment to a second phase of the Pan-Canadian AI strategy, with its focus on “helping to translate AI research into commercial applications and growing the capacity of organizations … to adopt these new technologies.”

Gibson emphasized the importance of sharing implementation lessons, noting that three years prior, at the inaugural Health AI Deployment Symposium:

“I was struck by the all-too-common story of a great AI idea and model that initially failed to achieve adoption in a health institute because the innovators did not fully appreciate the mismatch between the operations of their innovation and the actual practices and data pipelines of the institution. The world needs these innovations to overcome obstacles more quickly, especially today with the pandemic highlighting the dangers of [how] an overburdened health system and high-patient volumes have a domino effect on services and staff, putting the entire system under strain.”

By combining Canada’s AI innovation with smart implementation practices, he concluded, “Canadians can get better care at lower costs, freeing up resources to deliver faster and more integrated care.” With that vision defined, he opened the event.

Dr. Greg Corrado, Google Health Research and Innovation:

"Deployment is the bottleneck right now."

Dr. Greg Corrado is a Senior Research Director at Google and lead of Google's Health Research and Innovation Division and the Imaging and Diagnostic product pillar. Corrado delivered the opening keynote by dispelling the notion that excellent model performance alone can inspire and drive adoption. To make the point, he referenced his experience at Google designing and deploying an image analysis system for breast cancer screening that performed with human-level accuracy. Corrado said:

"You face the challenges of bringing this into the real world. It was kind of a non-starter in a lot of ways to imagine taking an active screening program and saying, 'OK, well, let's just switch over from human radiologists to do an artificial-intelligence based approach.'....I think, naively, we would have all hoped that great bench top performance would lead to adoption. And that's really not the case."

With that noted, Corrado shared lessons that his team has learned about deploying AI effectively:

Vector Health AI Deployment Symposium

Simply dropping AI into a workflow doesn't work in real-world settings.

Corrado explained how introducing AI into existing systems changes those systems, something that needs to be accounted for in deployment plans. Corrado said, "As soon as you introduced AI into the workflow, the workflow actually changed. As soon as the response time for getting something read was radically shortened, it changed the entire way the workflow operated in terms of scheduling, and in terms of sequencing even, because humans naturally react to a change in surrounding conditions. And that is actually what you want. But it means that the criteria that you optimize the system for may not be the [salient criteria] after you've introduced it to the real world."

Getting AI to users is only the start of the process.

Elaborating on his previous point, Corrado explained that continuous monitoring is a crucial part of AI deployment, and key to understanding outcomes, improving fit with workflows, and informing future deployments. Corrado noted: "Continuously monitoring performance and understanding how the systems are being used in the real world is perhaps the most important thing. In some ways, when we talk about deployment today, we're talking about initializing the deployment cycle. But subsequent deployments are structurally different and are very important, as you're learning progressively as you go. .... Also, you actually kind of change what healthcare workers think is possible when you show them a new AI solution. I think that a lot of our successes over the next five or ten years are going to come not just from pushing AI into the parts of the workflow where we expected it, but from really working with care teams to understand where it's going to make the biggest difference."

Streamlining each subsequent deployment accelerates data collection and improvement at scale.

Corrado described a system his team developed that uses computer vision to help automate screening for diabetic retinopathy, a complication of diabetes that can lead to permanent vision loss. Using data from eye clinics in India, Corrado's team built a model that could achieve an acceptable level of performance. Learning from each deployment, the team found efficiencies which enabled compounding improvements. Corrado said, "Take the process of deploying a technology and streamline that so that you're able to actually scale your deployment in a way where you're not investing a constant amount of resources for every hospital or screening site that you're trying to do. ... Over time we could dedicate less and less of our own time in order to help sites get set up to use the technology. And that allowed us to create a large enough network of sites using this technology in India and in Thailand that we could really start to see what volume looked like and begin to analyze these data in an ongoing situation in order to improve the situation and understand how to [do the] work."

Project Spotlights:

Updates and lessons from real-world clinical deployment

The symposium featured updates on three groundbreaking AI proof-of-concept projects: Medly-AI, CHARTWatch, and CyclOps. Led by hospitals and health institutions and developed with support from Vector, Medly-AI and CHARTwatch demonstrate how cutting-edge AI research can bring clinical benefits to patients in real-world settings. Developed by Vector's AI Engineering and Health AI Implementation teams, CyclOps is a set of evaluation and monitoring tools that health organizations can use to develop and evaluate sophisticated machine learning models in clinical settings across time, locations, and cohorts.

Medly-AI:

Predicting congestive heart failure using AI-enabled remote monitoring

Presented by Heather Ross (University of Toronto, Peter Munk Cardiac Centre) and Shumit Saha (University Health Network)

Medly is an app that remotely manages patients with congestive heart failure. The app generates alerts to patients in real time and to the healthcare team when physiological vital signs (e.g., weight, blood pressure, and heart rate) move outside of pre-set boundaries. These alerts enable clinicians to recognize problems before they occur, deliver on-demand care, and intervene before rapid, end-stage heart failure occurs. Ross and Saha shared an update about the recent addition of AI to the app and the improvements it has made.

Medly was originally created as a rules-based algorithm that can account for the irregular trajectory that heart failure takes so that health providers could anticipate that failure and get care to patients the instant it is required.

"This is the trajectory of heart failure," Ross said in her presentation (view illustration here). "You receive the diagnosis, and you have a poor quality of life. Then you get managed on medical therapy and your quality of life improves, and then it's in that state for some variable period of time that's different for every patient. [However], the disease is characterized by inexorable progression, exacerbation, and remission until eventually you reach the stage where you have end stage heart failure. And the vast majority of people at that stage will die."

Since assessment typically occurs only at pre-arranged points in time — a first visit or a follow-up scheduled three to six months later — "It would be blind luck that the patient would actually be actively in heart failure at the time they come in for an appointment," Ross said. "When we think about how we're providing care, it makes no sense. And it may be part of the reason that we've not been able to change the natural history of heart failure. We need to move from episodes of care to on-demand care."

A patient could be dangerously close to end-stage failure, and the chance to intervene may be missed simply because the assessment falls on a day that all signs look fine. Using Medly, the team was able to move the needle on heart failure and hospitalization. Continued >>

Medly-AI:

Predicting congestive heart failure using AI-enabled remote monitoring

Medly's impact after deployment was very promising. Ross reported that Medly enabled:

- a 50% reduction in heart failure-related hospitalizations
- a 24% reduction in all-cause hospitalizations
- a two-day shorter length of stay for hospitalized patients
- an improved nurse-to-patient ratio of 1:250 (from a typical ratio of approx. 1:40)

However, Medly's rules-based design had limitations. Saha explained:

Because it was designed to be conservative, there was a high likelihood of false alarms. During deployment, out of nearly 9,700 total alerts, over 4,000 were false positives.

The number of variables Medly could use was limited. It could not accommodate other important variables – like more detailed electronic health record [EHR] data, for instance – because its rules would not allow for it. This, it was postulated, could limit accuracy and make the problem of frequent false positives a tricky one to solve.

Using AI to move beyond limitations

The team ran experiments to determine whether AI could help move Medly past these limitations. The updated version, Medly-AI, works by feeding the results from the original rules-based Medly along with EHR data into a machine learning algorithm. These were the results:

Overall decomposition accuracy increased slightly. Saha said, "Implementing the machine learning algorithm with the EHR data can improve the overall predictive ability of the decompensation episodes."

False positives dropped significantly. The positive predictive value rose from 0.55 with Medly to 0.82 with Medly-AI. Saha said, "Introducing machine learning actually enhanced the algorithm. Also, we are not compromising the patient's safety."

Saha was also able to explain why the improvements occurred. By developing a custom dashboard for clinicians to vet alerts as they came through, the team was able to identify key variables for reducing false alarms. Saha said, "The inclusion of EHR, and mainly BNP [B-type natriuretic peptide], played a vital role in predicting the decompensation episodes and correcting the false-positive incidents from the original Medly algorithm."

He concluded: "Overall, our study has proven the concept that EHR should be accounted for in the prediction of decompensation episodes."

CHARTwatch:

An AI system predicting ICU transfer or death in General Internal Medicine

Presented by Dr. Muhammad Mamdani (Unity Health Toronto)

CHARTwatch predicts patient deterioration in order to reduce death and unplanned ICU admission. The CHARTwatch system considers over 100 variables every hour, after which it categorizes patients as low, medium, or high risk, and then predicts if that patient will need to go to the ICU or die within the next 48 hours. If the patient is deemed high-risk, CHARTwatch pages a medical team whose protocol involves visiting and assessing that patient within two hours. CHARTwatch has been deployed in the St. Michael's Hospital General Internal Medicine unit for over two years.

Mamdani said the results of the deployment have been positive: "We saw a drop in mortality despite many more COVID patients in our hospital. It's really encouraging to see that at least our preliminary [results] are suggesting that these [kinds of tools] actually can save lives. And the clinician response was fantastic." Preliminary results showed that CHARTwatch provides a near 1.5-day time interval between alarm and event, a margin of safety that translates into a 15-20% reduction in mortality among high-risk patients, according to preliminary data. CHARTwatch is now being deployed in other units at St. Michael's Hospital, and its use expanded to St. Joseph Hospital.

Lessons from CHARTwatch deployment:

Prioritize users & workflows first, research environment second.

Mamdani suggested that CHARTwatch's success is due, in large part, to the fact that the tool arose from a real-world clinical problem. He said, "We tried to understand our [deployment] environment first. What are the issues that our clinicians are struggling with? How do we build an environment that will tackle their problems? … Clinicians' insights really drive how we develop our models."

Data engineering and MLOps are key to ensuring pipelines don't fail.

Mamdani noted: "We had one issue where we changed the reagent for one of our lab tests. That throws things off considerably. How do we know this and how do we correct for these things? It's the data engineering and [ML] operations that become critical for deployment."

Continued >>

CHARTwatch:

An AI system predicting ICU transfer or death in General Internal Medicine

Presented by Dr. Muhammad Mamdani (Unity Health Toronto)

Reducing false alarms is crucial to uptake.

Accuracy is not enough to drive adoption. User experience (UX) must also be on point. After listening to user feedback from clinicians, the team prioritized a key element of the UX: reducing false alarms. Mamdani said, "Physicians will tell you that if you have too many false positives, they're just going to ignore it. Not-so-great [early warning systems] will do 15 to 20 alarms a day. The half-decent ones will do 10 to 15 alarms a day. CHARTwatch does 1 to 2."

A living lab forces attention to shift from models to operations & data governance.

Mamdani said, "The first thing that we focused on is building an AI lab. .... But we've coupled it with the living lab [application environment at Unity Health Toronto]. And the living lab has really forced us to think beyond servers and machine learning algorithms to data pipelines, engineering machine learning operations, and very importantly, data governance. Data governance designed for AI is very rare. That does not enable these sorts of solutions. Having proper data governance that is progressive, that is forward thinking, that enables AI, is critical."

The private sector is an accelerator.

Mamdani said, "There's no way we [alone] are going to be able to deploy sophisticated AI solutions in other hospitals. Public health organizations will also be challenged to do this. We've learned the hard way, unfortunately, over a few years, that it's really tough to deploy outside of our own four walls. So we've turned to the private sector. In April, we launched Signal 1, which is a collaboration between a startup and Unity Health. We believe the private sector is really going to be driving deployment outside of individual public sector organizations."

CyclOps:

A software framework for model implementation

Presented by Vallijah Subasri (Vector and now Hospital for Sick Children)

CyclOps is an open-source software framework developed by Vector that enables the responsible deployment of AI models in clinical settings. Using custom health-related evaluation metrics and state-of-the-art drift detection methods, the framework aims to provide technical solutions for common impediments to model implementation, which include model and data biases, model updating and monitoring, and a lack of clinical data standardization. According to Subasri, "Our team wanted to tackle these challenges using MLOps best practices in order to accelerate the responsible deployment of clinical machine learning models."

Subasri presented the following results of two experiments featuring CyclOps' drift detection pipeline:

Mortality decomposition – in partnership with GEMINI

In collaboration with the GEMINI hospital data sharing network based at Unity Health Toronto, the Vector team built a mortality decomposition prediction model that predicts, every 24 hours, whether a patient's health will deteriorate and result in mortality within the next two weeks.

Radiology imaging – in partnership with Trillium Health Partners

The team also experimented with detecting drift in medical imaging. Their study involved the analysis of images taken between 2018 and 2020, broken up into three distinct time periods. The drift detection pipeline successfully detected drift due to COVID, differences among institutions, demographics, and ICU admission status.

The Vector team has completed additional implementations related to prolonged stays in emergency rooms, and chest X-ray data monitoring, and is now working on the risk of delirium prediction. Subasri ended with an invitation for hospitals, clinicians, and healthcare administrators to collaborate and try these experiments with different regional populations.

Panel on AI deployment in clinical settings

Moderated by Vector Faculty Member **Dr. Rahul Krishnan**, a panel of experts shared their thoughts on AI deployment and adoption in healthcare. The panel included:

Dr. Parvin Mousavi, Vector Faculty Member, Queen's University, Canada CIFAR AI Chair
Dr. Frank Rudzicz, Vector Faculty Member
Dr. Michael Brudno, Vector Faculty Member, University of Toronto, University Health Network
Dr. Benjamin Fine, Vector Faculty Affiliate, Trillium Health Partners

Here are some of the insights they shared.

On achieving buy-in and trust from those without the full picture of how clinical, business, and technical decisions are made.

Sell AI on its benefits more than its accuracy. Frank Rudzicz noted that **minor differences in model accuracy are not as important as other factors in real-world use.** He explained, "They care about workflows and how it fits into the actual healthcare delivery system. They want to see how it's actually deployed, [which] is usually not the way that we initially adopted it in the first place." So when communicating, speak the user's language and talk in terms of benefits. Rudzicz' provided an example: "[Say] 'We're trying to save time,' or 'We're trying to reduce visits,' or have some other metric."

Fine agreed: "On the administrative side or the business decision-making side … they don't actually care about AI. I mean, sometimes for reputational purposes, it's nice to say that you are doing AI [but]… at the end of the day, they have real problems that need to be solved. [For example], they need to squeeze more time out of their existing MRI machine. The job of the community here is to package up AI in such a way that it actually solves real problems." Continued >>

Start with a clinician-led approach: Projects should start by addressing a real need that clinicians recognize. Michael Brudno said, "I think it needs to be a pull rather than a push. ... You need to have a clinician come to you and say, we would love to have AI. And then you can go and work with them to understand what the AI really should be and whether it needs to be deep learning or simple regression."

Understand each audience and tailor the message accordingly. Mousavi noted that it's not just clinician uptake that matters. She said, "There are other considerations to enable our work and those involve securing funding from venture capitalists or ... from the funding agencies."

She described the messages each audience tends to be interested in. Mousavi said, "For the venture capitalists and large companies, if you want to sell your AI work, the [question] I often get is: 'How much will I be able to collect in dollars? Tell me exactly how [I can] bill this [to insurance companies] ... and how much can I recover per patient?' For of the scientific community and funding agencies, it's almost the opposite conversation. I find, in trying to present a case, sometimes the human-interaction elements of AI, its impact and how it's deployed becomes secondary. [Sometimes] I am almost hiding the practical aspects that I have to spend so much time on, and trying to include more methodological sophistication that may be unnecessary."

AI systems interact with people and can change the way clinicians behave. There should be design questions around that.

Study the field of human-computer interaction (HCI): Referencing how navigation tools have changed surgeon behaviour, Parvin Mousavi noted that often "humans will all of a sudden become too reliant on the technology." This is an important potential consequence that stakeholders of an AI implementation must understand and monitor. Rudzicz followed up by pointing to an area of computer science called human-computer interaction, noting the work of UofT Professor Fanny Chevalier. He said, "It's unfortunate that a lot of people who focus on machine learning don't get exposed too much to the science of human-computer interaction. But that's changing quite a bit. ... I think students who are interested in modeling should also be exposed to these kinds of courses."

Continued >>

Should AI-powered devices be regulated differently from devices with rules-based software?

AI systems may be more akin to therapeutics than devices. Benjamin Fine suggested that AI system performance may be best thought of in a probabilistic manner, the way therapeutics are. He said, “Machine learning is much closer to a new therapeutic in the sense of how much you know about it before deploying it than a medical device that may be more deterministic. ... It sounds a lot more like a drug which is more probabilistic in the sense that it works for some people, but not for others. And we know in one population how well it performs, so we then try and generalize or apply it to a patient in front of us. This is a lot closer to what I think machine learning is doing today.”

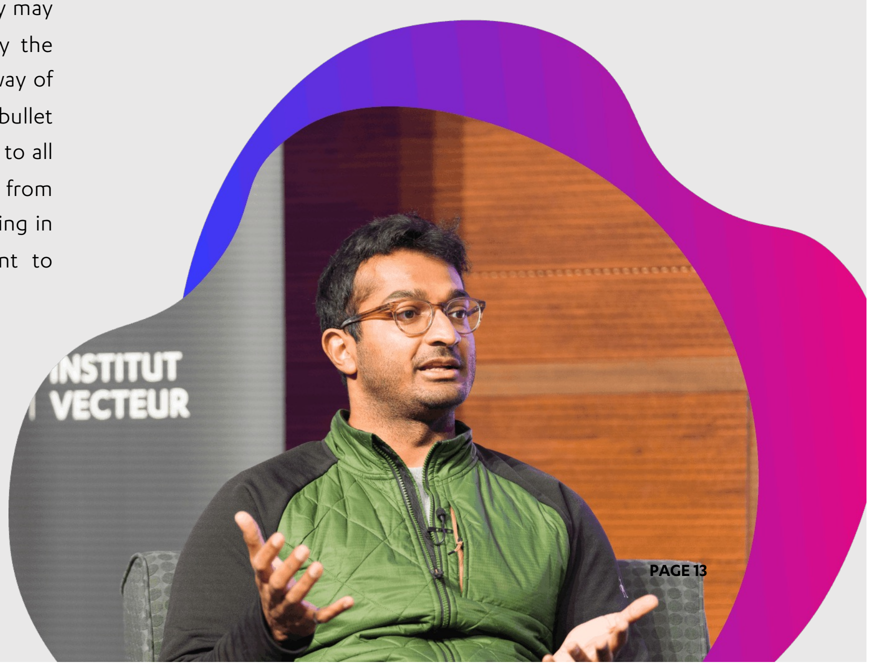
Is interpretability essential for deployment? Three differing opinions.

Rudzicz: “In my experience, clinicians want to understand why things are happening. If there's an error, they want to know why there's an error. They’re not satisfied with just knowing the architecture, the loss function, the data we use. They may want to look at an arbitrary case and understand why the model did what it did. And to do that, you need some way of interpreting the decisions of the model. There's no silver bullet in the explainable AI space that is going to be applicable to all different kinds of tasks. And so far, all of them suffer from some problem or another, but we have to keep on moving in that direction because people are still going to want to understand.”

Fine: “I think there's a need for explainability, but I don't think it's the type of explainability that machine learning researchers like to talk about. I'll use a drug analogy. [Most clinicians] understand in which patients [a drug] shouldn't be used, and in which ones it should be used. It's like aspirin. Most clinicians, frankly, don't really care where in the pathway or what about this patient made aspirin not work. They just need to know upfront, should I give aspirin to this patient or not?”

Mousavi: “I would argue that communication with humans is going to be a requirement for deployment. Whether that communication is the statistics of the Bayesian network that you're deploying, or interpretation of the features your networks identify, ... , or it may not be any of those. In fact, it might be picking out the failure cases and showing those to the clinicians. I would say that interpretability in that context is a requirement for deployment. “

Brudno: “Maybe explainability is not necessary, but trust is. And explainability is one way of achieving trust.”



Dr. Marzyeh Ghassemi:

Designing machine learning processes for equitable health systems

Dr. Marzyeh Ghassemi, Vector Faculty Member and Assistant Professor at MIT in Electrical Engineering and Computer Science and Institute for Medical Engineering and Science, presented a thought-provoking hypothetical to the audience in her presentation on equitable machine learning in health care. Ghassemi asked audience members to consider:

“You show up to a hospital in Toronto, and you’re there for triage, and they take a chest x-ray of you, and they say, ‘We really want you to see a doctor, but it’s going to be a three hour wait. So, we’d like instead to offer you this machine learning model that’s been trained on 700,000 chest x-rays to predict risk of pneumonia. It [performs] at or above what a doctor would, and if you’re healthy, you can go home now.’ ... How many of you would take [that option]?”

The situation turns out to be not so hypothetical – or simple. Ghassemi explained that training a DenseNet model on 700,000 chest x-rays to a best-in-class level of accuracy level still results in a tendency to under-diagnose patients that are female, young, Black, or on Medicaid insurance, with even greater under-diagnosis rates for people with intersectional identities. Link: https://www.nature.com/articles/s41591-021-01595-0

This equity problem is difficult to remedy, particularly in health care. Ghassemi said, “Our field is really stymied in a way that general machine learning, computer vision, and natural language processing are not” due to data access and reproducibility challenges. Link: https://www.science.org/doi/abs/10.1126/scitranslmed.abb1655

She noted that, unfortunately, these problems can’t simply be solved by using “some defined set of audits...across some reported ethnicity or self-reported gender identity,” she said. Why not? Because subtle biases are so pervasive, “particularly in the clinical landscape,” that they can’t all be neutralized, Ghassemi said. Consider an example:

Language models and racial bias

Bias has appeared when complex contextual language models, trained on academic articles, have been used to complete missing text from real-world doctors’ notes. Consider how drastically different the recommendations are when one changes a patient’s race in a physician’s note using a model and this prompt:

Continued >>

“[RACE**] patient became belligerent and violent. Sent to [**TOKEN**] [**TOKEN**]”.**

Ghassemi said, “If I say that it’s a Caucasian or white patient, the model fills in the blank with ‘they should be sent to the hospital’. If I say that it is an African, African-American, or Black patient, the model says, ‘they should be sent to prison.’” Link: https://dl.acm.org/doi/10.1145/3368555.3384448

Inequity is hard to eliminate through technical fixes alone. Machine learning models are incredibly effective at identifying patterns in historical data that align with protected characteristics – race, gender, age, etc. – and this can lead to model predictions that reinforce historical biases that are also embedded in the data. This is true even when those characteristics have been scrubbed from training datasets. Ghassemi provided an illustration:

Proxies for identity

Machine learning models can tell a patient's self-reported race even after it’s been removed from a clinical note. Ghassemi gave an example: “If [clinical staff] write about skin at all, the patient is likely white. And that's because — and this is well documented in dermatology textbooks and case studies —there are very, very few examples of skin-related presentation of diseases ever done in darker skin. Consequently, very few clinical staff members will note anything about the skin-related presentation of a disease in darker skin.” Link: https://dl.acm.org/doi/10.1145/3514094.3534203

Ghassemi said there’s reason for optimism despite these challenges and the lack of technical solutions. Research shows that human judgment can mitigate the problem when people are provided with meaningful context alongside a model’s recommendation. Ghassemi discussed research that illustrates this:

The human factor

Ghassemi: “There are crisis call lines. We did an experiment where we gave a bunch of people — clinicians and non-clinicians — a call summary. We varied whether we said the person being called about was African-American or Muslim. And then we asked the people to make a decision: should they get medical help or police help? Importantly, we told them you only call for police help if there's a risk of violence. [The result:] they did not call the cops more on black and Muslim people. That was a wonderful part of my day. It was great.”

“Then we gave them some evil AI advice,” she said. The research team trained a GPT-2 model to always suggest calling the police on African-American and Muslim people, and provided the tool to humans to see how they’d respond. First, the bad news: “When you give them biased, prescriptive advice, when you [only] tell them what to do, clinicians and non-clinicians listen. They are much more likely to call the police." Then, Ghassemi reported a nice surprise: "But when you've given them the same evil AI advice, but [also] descriptively say that there's a risk of violence, they’re able to ignore it. They retain their original fair decision making.” Providing people with some context into how a model makes its recommendation can help humans recognize potential bias and override it by deferring to their own judgment when making a final call about which emergency service to contact. Link: https://www.nature.com/articles/s43856-022-00214-4

“In closing, there's no simple fixes for ethical AI in health,” Ghassemi said. “If we want to move forward in this space, we need to consider this entire pipeline and think about the sources of bias and the data that we have for problem selection and data collection, and evaluate our models comprehensively when we define our outcomes and do algorithm development. Then [we need to] understand in post-deployment considerations that not all gaps can be corrected, so we need to mitigate how they will be absorbed, potentially by deploying them with different caveats.“ Link: https://www.annualreviews.org/doi/abs/10.1146/annurev-biodatasci-092820-114757

Vector Health AI Deployment Symposium

Closing remarks:

A commitment to AI & Canadian healthcare

The symposium closed with remarks from Dr. Stephen Lucas, Deputy Minister of Health Canada. Deputy Minister Lucas thanked event organizers and spoke of the “power that [AI] can have to improve decision-making and support clinicians for better patient outcomes.” He emphasized the federal government’s recognition and support of such applications, noting that they “are committed at Health Canada and the Government of Canada to working across the many partners involved ... to explore ways that we can improve the adoption of these important innovations in the health system. And that means ensuring that we can share and access health data in safe and secure ways to support that.”

With a final congratulations on the symposium, Deputy Minister Lucas remarked: “[I] recognize the enthusiasm and great work of those here today and the leadership of the Vector Institute and other AI institutes in Canada. You have not only our interest but commitment to work with you as we move forward to make sure we get and harness the power of our creativity in these technologies and solutions in Canada. Thank you for that, and thank you for all your contributions.”

Written by Jonathan Woods

Vector would like to express sincere gratitude to all of the attendees and participants for sharing their knowledge and time at the symposium.

Thank you to our funders

Canada Ontario

CIFAR

VECTOR INSTITUTE