Virtual Meeting
05/16/22 – 05/17/22

MEETING SUMMARY

Machine Learning and Health Outcomes in Cancer Care Delivery Research

NIH NATIONAL CANCER INSTITUTE

TABLE OF CONTENTS

BACKGROUND

The National Cancer Institute (NCI) conducted a virtual workshop focused on research in machine learning, cancer care delivery, and health outcomes. The goals were to highlight current cancer-focused research in health outcome risk prediction, symptom trajectories, natural language processing methods, and point-of-care decision support while identifying opportunities in this emerging area of cancer-focused research.

WELCOME AND INTRODUCTION

Katrina Goddard, Ph.D.
Division Director, Division of Cancer Control and Population Sciences (DCCPS), NCI

Roxanne E. Jensen, Ph.D.
Outcomes Research Branch, Healthcare Delivery Research Program, DCCPS, NCI

Dr. Goddard welcomed participants to the meeting and thanked the planning committee and meeting coordinators for their team effort inside and across divisions. She conveyed her enthusiasm and support for the future use of machine learning and natural language processing methods in cancer and healthcare delivery research. Dr. Goddard incorporated some of these methods into her own investigative work and understands its potential power to advance science. The field has grown rapidly over the past five years with an exponential increase in research grant applications. Approaches include methods to address a range of complex quality-of-care topics, such as early cancer detection and point-of-care clinical assistance to support symptom management. NCI is committed to backing research in this area to gain understanding of key research opportunities, needs, and challenges.

Dr. Jensen noted that healthcare delivery research is expanding the role of technology while repurposing data for research. Using a variety of sources, including healthcare claims and electronic health records (EHRs), the field is evolving to help cancer care professionals expand data use and capture information more efficiently. Studies in machine learning can help reduce the burden of patients and providers. Thus, this workshop features current cancer-focused studies in this area, while providing a platform for ideas and feedback from researchers, as well as individuals outside of cancer care.

Dr. Jensen introduced Ms. Abby Sears, chief executive officer (CEO) of the Oregon Community Health Information Network (OCHIN), a national nonprofit learning network for health equity research and innovation, who was recently selected to lead research for the new Artificial Intelligence/Machine Learning Consortium to Advance Health Equity and Researcher Diversity (AIM-AHEAD) program at the National Institutes of Health (NIH). OCHIN has grown since its inception in 2000 from a $7 million organization hosting EHRs for 20 health centers in Oregon to a $110 million organization that supports a growing national provider network with a diverse virtual workforce of more than 700 employees across 40 states.

Keynote Presentation: Driving Health Equity Through Representative AI and Inclusive Innovation

Abby Sears, M.H.A.
President and CEO, OCHIN

Ms. Sears began her talk by reflecting on health care from the perspective of her organization, which is based on the belief that people have the right to good health, and the collective community has a responsibility to achieve health equity. It is important to understand the populations who are uninsured, underinsured, unemployed, and underemployed and to bring their voices to the forefront of our delivery system. Recognition of deleterious influences, such as housing insecurity, food deserts, lack of transportation, lack of family support or caregivers, and underrepresentation in the community, on health is important. OCHIN believes it is important to describe the impact of these factors on health care and how technology and data can "[create] visibility for the invisible, as part of our delivery system." The company has envisioned health and equity since their start over 20 years ago. OCHIN acknowledges past racial oppression in the US healthcare delivery system, which concentrated on assisting commercial, insured populations. Her company has a commitment to ensuring a similar quality of care and technical assistance to underserved populations, with a focus on building trust, respectful representation, and partnership equality within the healthcare system.

OCHIN has been growing at a rate of 27% per year for over 12 years. In 2022, the company increased activity by 30% to touch 6 million active patients, with data on over 30 to 40 million patients. Their mission spans work in correctional facilities, school-based health centers, behavioral health centers, and substance-abuse treatment centers, where they deployed and host one of the largest EHR systems. They also provide support for dental clinics and public health departments with significant numbers of HIV/AIDS and LGBTQIA patients. The latest venture of OCHIN is assisting critical-access hospitals and their community partnerships to improve access to this aspect of the delivery system for uninsured patients or those on public assistance. Individuals receiving Medicaid or Medicare are often left behind by current healthcare models; however, they are the economic development engines of their neighborhoods and need independence in order to make community-wide contributions.

Her company is creating one virtual system, that has never existed before, with one master patient list or index, where all providers can document electronic health information. While there is one index, each organization has its own financial setup. The system represents the largest longitudinal and horizontal data set in the country for a unique patient set not obtaining care anywhere else. OCHIN is active in 46 states, touching over 21,000 different provider specialties. They move and share more data than any other organization in the country and allow data access for research purposes.

Patients accessing the system OCHIN created include women, children, and families, of whom 42% are below the federal poverty level. A third lack resources related to the social determinants of health (SDH), such as transportation, housing instability, and access to nutritious food, disproportionately impacting people of color in low-access rural and urban areas. One out of three speak English as a second language, and three out of five have chronic diseases. There are significant numbers of patients with mental health diagnoses as well, such as schizophrenia, depression, and anxiety, thus their care is complex.

OCHIN views their company as stewards of data to ensure that minority populations are not left behind through development of high-quality standards for representative data and future metrics. They are working upstream to collect information on SDH, touching 150 organizations, with more in the pipeline. In the next 18 months, they will have over 200 different organizations working on a shared system while providing technical assistance to 50 to 60 companies that are not part of EHR deployment. OCHIN currently touches 30% of patients seen through federally qualified health centers, providing opportunities around natural experimentation and best practices for innovation. The services provided by her organization help to offset costs when community health centers endure budget cuts as well.

OCHIN is also involved with cutting-edge innovation, such as:

- Active research projects through various organizations and vendors, with 60% member participation
- Peer-reviewed and transformational projects
- Telephonic care in rural communities that lack broadband access, including $43 million in Federal Communications Commission funding to improve broadband and remote patient assessment
- Cybersecurity and regional extension center partnerships

OCHIN was one of three organizations in the United States that worked with the Centers for Disease Control and Prevention (CDC) to build an electronic case reporting system and national framework during the pandemic to transfer data from 250 different public health agencies in 50 states. They were also instantaneously moving data on all 6 million patients, and subsequently found errors in their algorithms which impacted contact tracing, that they are working to improve. The company is partnering with the CDC on an advanced clinical research database with 8 million active patients. This project combines machine learning with screening to obtain early testing in target populations within the clinical environment.

OCHIN was recently awarded one of the domains related to NIH's AIM-AHEAD initiative to ensure patient equity and bring minority and underserved populations into decision-making positions related to healthcare policy and research. A powerful component is the fellowship program to engage more diverse researchers with the objective of improving access to care for marginalized and complex patient populations. Machine learning with database administration can help sort and define information to eliminate bias, reach vulnerable individuals, and provide culturally competent care.

Other challenges that exacerbate structural inequalities include provider/specialist access and caregiver burden, with 40% of those replying to a satisfaction survey stating that they are experiencing burnout. For example, California lost 30% of their providers, but care has increased approximately 20-fold. This situation is not sustainable, and OCHIN's artificial intelligence and machine learning systems are helping to alleviate complicated healthcare system needs, including those surrounding precision medicine.

However, it is important to note that it can be difficult to enhance health equity while not recreating structural inequality. Methods must be used that do not add costs to community clinics that may lack adaptive capacity and who serve underrepresented patients, including those on Medicaid. Researchers must recruit diverse cohorts and obtain data that can be trusted by vulnerable populations. Healthcare professions must develop solutions that address problems for all patients. We must all consider what quality care looks like to patients who face additional barriers, and how we can use artificial intelligence (AI) to incentivize care and drive better outcomes, transparency, and accountability.

SESSION 1 Landscape of NCI-Funded Research

Moderator

Ngoc Tran, M.D., M.S.
Surveillance Informatics Branch, Surveillance Research Program, DCCPS, NCI

Panelist

Roxanne E. Jensen, Ph.D.
Outcomes Research Branch, Healthcare Delivery Research Program, DCCPS, NCI

Dr. Jensen provided information on machine learning research trends in cancer care delivery, highlighting a portfolio analysis from NCI grants. This compilation was completed by Dr. Jensen and her colleague Dr. Bryan Kim. The objectives of this research were to use machine learning methods to help predict and analyze patient health outcomes with the assistance of expanded EHRs and data granularity. The growing digital space allows for unique analyses of patient-generated data with better techniques to predict risk and model complex patient trajectories.

Important questions related to this workshop and machine learning investigations include:

- What are the strengths of applying machine learning methods in this area versus "standard methods"?
- What are current limitations of machine learning in this area of research?
- How do we expect this area to grow in the future?
- What aspects of this area are specific to cancer care delivery?

Machine learning methods for healthcare delivery are taken from another field and adapted for use, thus these types of studies often focus on method validation such as natural language processing or generation. Other research focuses on the application at group or population levels to predict or characterize patients with higher rates of poor health outcomes. Lastly, there is also research that reviewed machine learning for the integration of clinical health information at the point of care (e.g., support for lung cancer screening).

They conducted iSearch and QVR database searches from 2016 through 2021 of all NCI grants containing terms such as *machine learning*, *deep learning*, and *artificial intelligence* in the title, abstract, or text. The objectives were to review the trends and NCI applications related to machine learning methods that focus on cancer care delivery and report health outcomes. The ultimate goals were to report on high-level programmatic details and examine application differences in research methods and clinical care applications. They identified 193 studies for potential inclusion and excluded 100 because these lacked healthcare delivery components or patient outcomes. Content coding variables included tumor site, age, cancer control continuum, study design, primary outcomes, machine learning methods, and application focus (research, methods, or clinical care).

Programmatic results show that 16% of the research was funded with overall applications steadily increasing since 2016. The majority of projects were small (20%), with some larger studies (14%) spanning 5 years or more; 15% were early-stage investigator. Regarding cancer type, 77% were hematologic, 50% breast, 46% lung, and 42% gastrointestinal. Across the cancer continuum, active treatment was the most common area of focus, along with screening and survivorship. Top outcomes included symptoms,

function, and adverse events (38%); healthcare utilization (34%); diagnosis (25%); and care quality metrics (15%). While machine learning methods were most likely to be funded, the majority of grants were research application (84.5%), with an increase in these since 2019. Approximately 30% were machine learning methods, and 35.5% were clinical care application. See the breakdown of study focus/design and research applications based on topic and outcome in the tables below:

Research Application – Spotlight on:	***Study Focus***	***Study Design***
Risk Prevention	Algorithm Development – 50%	Primary Data Collection – 24%
	Algorithm Application – 14%	Experimental Design – 4%
	Development & Application – 36%	
Symptom Trajectory	Algorithm Development – 37%	Primary Data Collection – 11%
	Algorithm Application – 14%	Experimental Design – 37%
	Development & Application – 49%	
Natural Language Processing	Algorithm Development – 49%	Primary Data Collection – 24%
	Algorithm Application – 18%	Experimental Design – 7%
	Development & Application – 33%	
Clinical Decision Support	Algorithm Development – 40%	Primary Data Collection – 55%
	Algorithm Application – 11%	Experimental Design – 30%
	Development & Application – 39%	

Research Application – Spotlight on:	***Cancer Screening***	***Diagnosis***	***Care Quality Metrics***	***Healthcare Utilization***	***Patient-Provider Communication***	***Symptoms, Functions, Adverse Events***	***Recurrence, Progression, Remission***	***Survival***
Risk Prevention	10%	22%	10%	24%	8%	24%	6%	11%
Symptom Trajectory	0%	5%	5%	47%	18%	66%	16%	18%
Natural Language Processing	9%	24%	15%	21%	20%	39%	6%	3%
Clinical Decision Support	15%	22%	19%	30%	22%	41%	19%	15%

In summary, NCI research grants related to machine learning are a small but rapidly growing area of research. Applications are likely in response to targeted funding announcements and often focus on screening and active treatment. There are few applications proposals that use established, validated machine learning models, but natural language processing is a popular method. Applications that focus on clinical care integration should carefully consider scope and expertise.

SESSION 2

Applying Machine Learning to Identify High-Risk Patients and Outcomes

Moderator

Andrew Freedman, Ph.D.
Clinical and Translational Epidemiology Branch, Epidemiology and Genomics Research Program, DCCPS, NCI

Panelists

Amrita Basu, Ph.D.
University of California, San Francisco

Chris Sander, Ph.D.
Dana-Farber Cancer Institute

Sherry-Ann Brown, M.D., Ph.D., FACC, FAHA
Medical College of Wisconsin

Shreyas Lakhtakia
Flatiron Health

Dr. Freedman moderated the second session on the application of machine learning (ML) to identify high-risk patients and predict health outcomes. The session included presentations from experts in the application of real-world ML, including identifying patients at high risk for cancer development and algorithm creation to predict severe treatment-related, adverse events in cancer patients. The biographies for each panelist were available in workshop materials or online at the workshop website at **https://events.cancer.gov/hdrp/machinelearning/speaker-bios**.

Shreyas Lakhtakia – Prospectively Using Machine Learning to Improve Patient Outcomes

Mr. Shreyas Lakhatakia provided information on the prospective use of ML to identify high-risk patients and improve patient outcomes. He stated that high-risk patients are important in cancer care, because nearly half of this treatment is provided in acute care settings, often beginning in the emergency department. Even more important is that more than half of these visits are preventable. ML has the potential to prospectively identify these high-risk patients in clinical practice settings to improve outcomes, including lower healthcare costs and increased access to care.

Unfortunately, only 2% of ML research articles focus on the prospective use of this technology to improve patient outcomes. Part of the problem may be the challenge of adapting ML frameworks to application in clinical practice. Testing of model performance is flawed and a lack of high-quality continuous data streams may create bias, or ML does not fit with user workflow.

Thus, Flatiron Health partnered with Huntsman Cancer Institute to design a framework for the prospective use of ML. The institute has a home care program to help acute cancer patients at risk by sending clinical staff to provide support at home and by phone. Thus, they hypothesize that using ML-enabled risk identification could preempt enrollment of patients in need, usually identified by providers. The process starts with defining a very specific outcome, then following typical ML workflows of model creation and evaluation. This analysis is then extended to evaluation in clinical settings with eventual deployment in the

clinician workflow. By starting with an outcome, the ML model can be built in pursuit of a defined clinical objective with a higher chance of success in actual practice.

The objective of the Huntsman@Home program was to reduce avoidable emergency department (ED) visits and acute care spending for at-risk patients. Enrollment in the program reduced ED visits by 45% and readmissions by 55%. This information provided the clinical objective to work back to the specific intervention for patients determined to be at risk, translating into the ML outcome the organization wanted to predict ED visits within 60 days. Typically, patients were recommended for the program if they were deemed to be at risk by their physician or at ED discharge.

Using ML, they identified these patients earlier and more efficiently than traditional methods. They trained the logistic regression model using structured data from the Huntsman Cancer Institute, controlling for factors such as race, gender, and ethnicity. Results from a prospective study were promising as the accuracy of real-world prediction of the probability of ED visits for the studied cohort within 60 days improved threefold. Feedback from providers showed that the technology was a good fit in the clinical workflow.

Dr. Chris Sander – Prediction of Pancreatic Cancer Risk from Real-World Clinical Records

Dr. Sander discussed a project on pancreatic cancer, which is a high-risk and aggressive form of the disease. Eighty percent of pancreatic cancer is identified at stage 3 or 4, when the cancer has advanced locally or metastasized, respectively. Therefore, the goal was to predict pancreatic cancer from electronic medical records and shift more patients to earlier detection to achieve better outcomes.

However, although the risk of pancreatic cancer increases with age, the actual incidence is rare – occurring in approximately 4% of the population. Thus, catching the disease is difficult, and routine screening would be cost prohibitive and ineffective. So ML methods were employed to solve this problem, using time trajectories of disease codes and prediction intervals for cancer after assessment. Using real-world longitudinal data on approximately 8 million patients from the National Patient Registry from Denmark and 1 million patients from the Cancer Registry in Boston, they explored over 2,000 International Classification of Diseases (ICD) codes with some modifications and used natural language processing to develop a model. The objective of the model was to create a practical tool to identify patients at elevated risk for pancreatic cancer to enroll in screening and disease prevention programs for early detection. Eventually they were able to determine high-risk predictors or grades of risk for pancreatic cancer over time, including male sex, type 2 diabetes mellitus, unspecified jaundice, hypercholesterolemia, and acute pancreatitis.

Dr. Amrita Basu – Predicting Immune-Related Adverse Events in Breast Cancer Patients

Dr. Basu addressed the dearth of new cancer drugs on the market today with excellent outcomes, such as immunotherapy. She stated that these medications have short- and long-term side effects, which may be intolerable to patients. The goal when developing a cancer agent is to achieve high efficacy, low side effect burden and duration, good quality of life, and no mortality. In traditional clinical trials, safety and tolerability are assessed at the end of treatment. Researchers now want to assess these outcomes during active therapy and prevent severe adverse events from cancer treatment.

Objectives include:

- Minimizing toxicity while maximizing efficacy.
- Quantifying the likelihood of developing a chronic condition.
- Understanding clinical symptom predictors that contribute to impaired quality of life.
- Understanding genetic susceptibility.

The adverse events currently under study are related to immunotherapy administered to patients across various cancer centers, and include chronic conditions such as thyroid dysfunction, colitis, adrenal insufficiency, and pneumonitis. Thus, Dr. Basu and her colleagues used machine learning to create decision support predicting drug toxicity, which can assist with potential treatment redirection. In addition, the group sought to understand the clinical predictors contributing to impaired quality of life.

Using data leveraged from the I-SPY clinical trial for breast cancer at 26 US sites (N > 2,000), a Bayesian adaptive model was created to test combinations of agents to prevent metastatic disease, balancing tolerability with efficacy. Patients were randomized to the control or experimental arm and receive experimental treatment for 12 weeks, followed by anthracycline for another 3 months. Looking at symptom clusters and severity grade associated with individual drugs, combined with baseline comorbidity, demographic, and disease information, Dr. Basu reviewed outcomes related to quality of life. Current data showed 23% of patients had at least one adverse event, with hypothyroidism being the most common. Computing symptom burden for hypothyroidism, they were able to identify two early symptoms – shortness-of-breath and rash – that were predictive of this chronic condition.

The researchers also collected patient-reported outcomes (n=137) for 24 months through an electronic platform, including adverse events and quality of life, with a 90% completion rate to date. Side effects which remained more constant included abdominal pain, cough, hot flashes, and insomnia, while joint pain appeared later in the course of treatment. Her group also is reviewing symptom changes from baseline, severity distributions, and persistence, along with symptoms that predict quality-of-life outcomes. Study limitations include limited sample size and patient and clinician bias in the data sets, such as under- or overreporting. However, they are incorporating a feedback tool to provide information on poor ML performance. Genetic susceptibility will be examined next.

Dr. Sherry-Ann Brown – Machine Learning in Cardio-Oncology

Dr. Brown discussed ML for risk stratification in cardio-oncology. To date, there is little information on which patients are a greater risk of developing adverse cardiac effects, such as cardiomyopathy, myocarditis, conduction disorders, or pericardial disease, from cancer therapy. ML can take vast amounts of patient data from various institutions to study this topic. Dr. Brown is part of a multicenter research group working together to produce ML algorithms in a long-term, cancer-survivor cohort (N=4,626), comprised of patients >18 years of age with available baseline cardiac imaging.

The group examined patients with a new cancer diagnosis who also developed a cardiovascular condition. They then correlated the most frequent cardiac diagnoses – cardiomegaly, cardiomyopathy, and heart failure – associated with particular cancer agents. ML was used to predict various adverse cardiovascular outcomes in this patient population for use by oncologists and cardiologists. The researchers are modifying data and technology for the algorithm and clinical decisions at the visual interface to impact prevention. Strengths of the study include subjects from multiple health systems, quantum computing and ML, and interdisciplinary teams, while limitations include data bias, heterogeneity, generalizability, and scalability.

Discussion

The panel was asked to reflect on the advantages of ML models versus past methods used for risk prediction. Dr. Brown replied that researchers have developed some risk scores for the development of cardiac disease after cancer treatment, but none have been validated and generalized, and the only ones available are for one or two types of cancer. The goal of her team is to develop a tool that is usable across the spectrum of cancer patients seen by cardio-oncology. ML can investigate patterns and predict outcomes from large data sets. Dr. Basu stated that since there are so many combinations of drugs/symptoms and adverse events, the scalability and granularity provided by ML to understand the factors underlying specific clinical presentations is important. Dr. Sander remarked that the advantage of ML is the ability to work with large, complicated, and heterogenous data sets, using multiple parameters. It can be difficult to analyze data from real-world systems using standard methodology. Mr. Lakhtakia

responded that there were three main advantages to using ML – data source integration, resources, and the ability to solve prediction problems that were previously impossible with traditional methods.

The next question posed to the group related to the major challenges encountered in the development of an ML model. Mr. Lakhtakia stated that there are biases related to historical patterns that may be replicated in the data. Data quality and determining what information to include can also be difficult. Real-world care may also change, making training models obsolete. Dr. Sander concurred that data quality and access are major issues, as is the protection of health information on computer and cloud systems. Dr. Brown stated that the institutions she works with have high-performance computers. She also remarked that models trained on only local data to assess cardiovascular risk in general populations do not perform as well as those trained on national data sets. Currently, there is no national data set examining the topic her group is investigating – cardiovascular risk from cancer medications. Her group is working to build a national registry starting with 25 US sites, using natural language processing to abstract data with ML to prospectively look at model outcomes. Dr. Basu mentioned that there is a lot of missing data, so her group is working to improve survey completion rates and technology access. In addition, clinical trials vary in methodology, with the collection of different symptom and quality-of-life measures.

The panel was also asked to discuss how they plan to put results from ML into practice. Dr. Brown responded that they are attempting to include diverse groups in their cohort, parsing out demographic information from the database to use in clinical practice. Her group is also pondering stakeholders and how the information might be useful in clinics, including clinical decision-making, app documentation, and electronic medical records (EMRs). Dr. Sander stated that the most important result from ML is integration into clinical practice, including analyses related to clinical trial design. Dr. Basu replied that they have an Outcomes Working Group which includes clinicians, site coordinators, and patient advocates. The goal is to take the data and intervene early, but they cannot implement clinical decision-making tools that generate unnecessary alerts, which often occur with traditional methods. Mr. Lakhtakia stated that ML should be in service to the patient, which means that decisions should be made to determine thresholds for high-risk and low-risk outcomes. He also stated that cross-functional teams are important.

Lastly, panelists provided opinions on future ML funding targets to improve oncological care. Dr. Basu stated that funding should be provided to improve data quality and quantity. Dr. Sander replied that healthcare consortia should be developed to share information. Mr. Lakhtakia remarked that money should be provided to help professionals implement tools developed by ML.

SESSION 3

Machine Learning Approaches to Improve Cancer Patient Symptom Monitoring and Health Outcomes

Moderators

Ashley Wilder Smith, Ph.D., M.P.H.
Outcomes Research Branch, Healthcare Delivery Research Program, DCCPS, NCI

Dana L. Wolff-Hughes, Ph.D.
Risk Factor Assessment Branch, Epidemiology and Genomics Research Program, DCCPS, NCI

Panelists

Chris Sidey-Gibbons, Ph.D., M.D.
Anderson Cancer Center

Dana Rollison, Ph.D.
Moffitt Cancer Center

I-Chan Huang, Ph.D.
St. Jude Children's Research Hospital

Jukka-Pekka "JP" Onnela, D.SC.
Harvard T.H. Chan School of Public Health

Dr. Dana Rollison – Machine Learning Approaches to Improve Cancer Patient Symptom Monitoring and Health Outcomes: The Importance of Enterprise-Scale Data Capture

Dr. Rollison discussed opportunities for ML to advance patient-reported outcomes (PROs) in oncology, such as baseline information, follow-up symptoms, and quality of life. Use of ML can be applied both clinically as well as in research development to produce algorithms that advance and promote future oncology care. Opportunities include extracting information from unstructured text, analysis of streaming data from mobile devices and across scales, hypothesis generation through pattern recognition, and adaptive learning of predictive models over time.

PRO measurement has immediate clinical utility, but it quickly and efficiently generates very large and heterogenous sample sizes for research and discovery as well. Thus, the data produces more generalizable algorithms and is particularly useful for minority populations. Also, prospectively monitoring algorithms in clinical settings over time is necessary to advance technology.

Dr. Rollison discussed lessons learned from the Moffit Cancer Center. Upon admission to the center, clinicians establish a patient baseline to compare trajectories during care. Patients complete a clinical intake form which electronically captures PROs, such as disease symptomatology, treatment side effects, and quality of life. There are core modules with standardized questions posed to all cancer patients, as well as disease-specific supplements which allow for flexibility. Discrete responses are stored in a data warehouse, while physicians receive a summary report through the EMRs.

Over the past seven years, they have had more than 20,000 unique patients, with 76% to 82% registering for the patient portal and 70% to 80% completing the patient questionnaire (PQ). Those who complete the PQ were more likely to be female, non-White, non-Hispanic, and younger. Seventy-nine percent of breast cancer patients completed the PQ, compared with 58% of patients with a diagnosis of lung cancer.

Those in later cancer stages also were less likely to enter information, which should be considered when interpreting the data. The center is currently targeting interventions to improve completion rates.

In 2016, they initiated the 2016 Edmonton Symptom Assessment Scale (ESAS) across multiple clinics. In 2021, they added an additional dimension on suicidal ideation. They graphically review the trajectory of these symptoms over time within the EMRs to determine areas for physician intervention. They currently have 65,000 ESAS records on 16,000 patients, providing large amounts of information for ML. The radiation oncology clinic was an early adopter of PRO-driven outcomes research and has already published on the topic. Given the high symptom burden following radiation treatment, this information can assist with radiation planning to reduce treatment toxicity. For example, they received a US Department of Defense grant using ML to predict PROs in prostate cancer patients to reduce toxicity and improve quality of life. Dr. Heather Jim and Dr. Issam El Naqa are developing models to determine patient outcomes in off-target areas, and whether they should be managed differently to reduce reported symptoms. While these systems are costly and there is no reimbursement path, showing the value behind this data can be useful.

Dr. Chris Sidey-Gibbons – Opportunities at the Intersection of Patient-Reported Data and Machine Learning

Dr. Sidey-Gibbons discussed the value of patient-reported outcome measures (PROMs) as an effective intervention in oncology care. He focused on opportunities to improve PRO using computational techniques such as ML and employing data to predict outcomes in breast and ovarian cancer. PROs and PROMs are reported directly by the patient without interpretation and include quality-of-life, symptom burden, and mental and physical health. This information is robust, reusable, and cost-effective in clinical practice. Providers also are finding that this information has clinical utility. A large Cochrane review of 116 randomized controlled oncology trials found that PROMs improve patient-provider communication, diagnosis, and patient quality of life. Moreover, PROs can even improve patient survival in certain settings.

PROs are effective because they reduce the time between treatment-related toxicity and clinician recognition, which could avert missed therapy. Using mobile-PRO symptom reports allows greater provider responsiveness to patients after the initiation of systemic therapy. However, responding to PROMs places a burden on patients at a time when they are sick. Thus, Dr. Sidey-Gibbons and his colleagues pursued anticipatory modeling to determine patient outcomes prior to treatment. They used computational techniques, such as computerized adaptive testing, to make PRO data easier to collect and more accurate. His team also found that providing personalized written feedback improves patient engagement and questionnaire completion. Natural language processing provides more accurate constructs and inferential psychometrics may be possible using digital footprints to successfully predict psychological traits from online behavior.

For example, his group used ML to predict satisfaction with breast reconstruction one to two years post-mastectomy. With data from the clinical portal on PROs, they applied ML models to predict the preferred method – autologous (flap) reconstruction or implants. Although standard studies found autologous flap was usually preferred, their data showed that most women preferred implants, with good performance (area under the receiver operating characteristic [AUROC] curve > 0.85 at two years). Another study awaiting scientific review predicted end of life for women with recurrent ovarian cancer. Patients with ovarian cancer often experience aggressive care, even up until death, which is contrary to national guidelines. This issue most likely results from the inability of clinicians to detect the clinical transition marking treatment futility. Related research from EMR-based studies also had poor sensitivity to detect short-term patient mortality. PRO data analyzed with ML produced algorithms that performed particularly well, with greater sensitivity in predicting patient death (AUROC = .76).

Dr. Sidey-Gibbons stated that predicting outcomes using ML could optimize care delivery at the individual level, while PRO assessment can be made shorter, more accurate, and more acceptable to patients with computational techniques.

Dr. I-Chan Huang – Artificial Intelligence-Driven Risk Prediction Based on Patient-Generated Health Data for Childhood Cancer Survivors

Dr. Huang stated that the five-year survival rate for pediatric cancer has improved, with an estimated 500,000 childhood cancer survivors in the United States. However, there are late effects related to cancer therapies, such as secondary cancers, chronic health conditions, neuro-psychological and physical performance deficits, and premature mortality. The goal of his work is to improve symptom research and facilitate survivorship care.

In a study of 735 adult survivors who participated in both the Chronic Health Conditions and St. Jude LIFE studies (1994–2018), they evaluated the prognostic value of baseline symptom clusters (T1) with cluster changes (T1–T2) on future clinically determined chronic health conditions (CHCs). Global results (respiratory and neurologic) showed that those with moderate physical and emotional symptom burden at T1 had a consistently moderate symptom burden at T2 (OR 1.42 [95% CI 1.03–1.97]), while those with high symptom burden at T1 had worsened or consistently high symptom burden (OR 1.63 [95% CI 1.18–2.26]) at T2. Greater risk was seen with individual CHC groups for the progression of neurologic symptoms.

St. Jude Children's Hospital uses the Momentary Symptom Assessment or Symptom Assessment Management (SAM) program using the mHealth for Childhood Cancer Survivors initiative. This patient-generated data, entered through smartphones and the DatStat platform, create risk prediction models for quality of life and clinical outcomes. Results are integrated into a personalized risk prediction through EMRs and augmented with decision support tools, which facilitate survivorship care.

Artificial intelligence using natural language processing (NLP) architecture helps to capture and convert conversation-based, unstructured PROs. The researchers then validate NLP/ML algorithms for analyzing this unstructured symptom data using expert-labeled symptom attributes. Bidirectional Encoder Representations from Transformers (BERT) also assists with word embedding and validation. The area under the receiver operating characteristic curve for identifying features of unstructured symptom data was .92 for cognitive symptoms related to pain interference and .95 for fatigue, using BERT. These results were superior compared to other NLP/ML models. Future use of AI-enabled PROs is aimed at identifying adverse events and facilitating decision-making to decrease clinical workload and improve the quality of patient care.

Dr. Jukka-Pekka "JP" Onnela – Digital Phenotyping for Cancer Patient Symptom Monitoring and Health Outcomes

Dr. Onnela first used cell phone data to study social behavior in 2004. Using anonymized metadata from detailed call records, he published an article on the structure and strength of mobile communication networks. Digital technologies provide footprints that are available for analysis, and there is great potential for its use since 6.6 billion people worldwide use smartphones. The goal is to build phenotypes or outcomes for patients.

However, digital phenotyping in social, behavioral, and cognitive domains present a special challenge because of their temporal nature and context dependence. There needs to be a scalable method to measure phenotypes objectively, or the "moment-by-moment quantification of the individual-level human phenotype in situ using data from personal digital devices." The advantages of using this type of data include the ability to capture the heterogeneity of illness presentation at the population level and the low burden of follow-up to capture trajectories over time.

The Digital Phenotyping Project required infrastructure, such as a custom, scalable platform for smartphone-based digital phenotyping. However, since large amounts of data are generated, statistical ML methods or AI were needed to interpret the information. There are several active studies in psychiatry, neurology, surgery, and other fields with the goal of systematizing data collection and analysis in smartphone-based digital phenotyping.

Their raw data collection platform, Beiwe, was developed in 2013, with front-end users (Android iOS applications) and the back-end Amazon Web Services cloud computing on a scalable and globally deployable platform. The Forest data analytics library incorporates an increasing number of methods, such as Python or integration with Tableau. Workflow on the Beiwe research platforms includes the creation and management of studies using a web-based study portal, app downloaded by subjects, collection and storage of encrypted raw data, modeling and analyzing data, and sharing the results.

There are two paths to obtain data, including prepackaged data summaries through software development kits (SDKs) available through companies such as Apple and Google and working with raw data. The problem with proprietary software is the use of closed algorithms, so content is unknown. Algorithms also evolve very quickly and require rapid modification to maintain validity. Since there is quite a bit of data captured (200 megabytes to 1 gigabyte of raw data per person/month), the analytics are complex, but AI may not always function correctly in the real world. Thus, combining data with various methods may be necessary. The goal is to work with interpretable data combined with transparent modeling, which can be achieved with ongoing information captured through the smartphone.

Discussion

Panelists were asked to reflect on the strengths of ML for symptom management and health outcomes. Dr. Onnela replied that ML allows researchers to learn about the data as its being collected. Dr. Rollison stated that she is excited about the ability to synthesize data from a variety of scales to improve precision. There are a variety of input methods, and physicians have limited time to spend with patients, so ML can assist with data management and interpretation. Dr. Sidey-Gibbons remarked that ML can be used to complete tasks that humans have trouble performing, although determining the gold standard is sometimes difficult. It can also be used for more low-stakes variables to obtain data not previously available. Dr. Huang commented that it depends on the type of data and the purpose of collection. ML can be useful for pattern analyzation and creation, as well as NLP and unstructured data interpretation.

Dr. Wolff-Hughes then asked about the concerns or challenges related to AI machine learning to meet the unique demands of vulnerable and minority communities. Dr. Rollison replied that ML provides generalizability if the algorithm was trained on a heterogenous population. Researchers do not know the variability when there are limited demographic differences within an analyzed group. Dr. Huang concurred that equity disparity is an issue, and he worried about equal access to technology to provide PRO data. Rural, aging, or low socioeconomic status populations may not be able to access these devices, which can impact data collection and thus results. Dr. Sidey-Gibbons emphasized the importance of scalable data collection methods from diverse populations on PROs, along with investment in community access to smartphones. However, Dr. Onnela commented that there are differences between minority and White populations in the type of device used, with iOS available in many academic research hospitals, which creates bias. He also stated that more foundational development on the deep-learning theory underlying these models must be obtained.

Dr. Rollison was then asked how her group determined the causal relationship between radiation and toxicity. She stated that the study objective was to demonstrate that ongoing symptom capture at each visit in the radiation oncology clinic could provide predictor variables but can also be used to model outcomes to advance precision medicine.

Next, panelists were asked to consider best practices for the theoretical underpinning of algorithm development, including the variables they choose to include. Dr. Sidey-Gibbons stated that methods such as excluding demographic information are important so that bias is not learned and replicated within the model. After model development, performance can be assessed in different groups. There are two ways to decide on variables, including statistically to maximize performance for sensitivity or specificity, and collaboration with end users.

Dr. Wolff-Hughes also asked panelists to deliberate next steps for the application of prediction models based on PROs to real-world, pediatric cancer survivorship care. Dr. Huang replied that they incorporated their model into EMR, but he also believes this research needs to be disseminated, with tracing of specific areas of integration. In follow-up, Dr. Onnela commented on how this task might be accomplished, along with the associated opportunities and challenges that might be unique to this type of information. Dr. Onnela stated that he believes it may be too soon for EMR integration, secondary to model accuracy. His group tries to validate their algorithms, whenever they are able, to identify a meaningful standard. For example, imputation can be used when there is missing data, particularly since population-level compliance with wearable devices may be low and inconsistent. Variations in data type and completeness also require careful exploration of collected information, and thus more work needs to be done at the population level in the next five to six years.

Panelists addressed how PRO data have been used to assess drug effectiveness in real-world therapy. Dr. Rollison stated that this type of research is an emerging area, and they are reviewing it at the Moffitt Cancer Center. While PROs are important and should be collected on a greater scale, a challenge is conducting across institutions and lack of standardization in data and methods. Dr. Onnela remarked that they are using smartphones as part of clinical trials, as both a primary and exploratory endpoint. For example, in amyotrophic lateral sclerosis (ALS) research, they collect information on activities of daily living, and thus they are using this technology to directly measure clinically important symptoms related to disease outcomes. In real-world settings, ML models have been used in clinical trials.

Dr. Huang commented on what can be learned from systems that gather large amounts of intermittent global positioning systems (GPS) data on variables, such as the social determinants of health. He stated that they use GPS data to understand risk factors such as environmental exposure, healthcare utilization, and activity levels.

In conclusion, panelists were asked how they think ML can be used for symptom management in cancer patients and future evolution of the field. Dr. Rollison replied that standardization and access to robust data across centers are necessary objectives to advance the prediction of PROs and other outcomes. Dr. Sidey-Gibbons stated that computational approaches will improve and transform symptom monitoring. Using data science capabilities to formulate questions at a better cadence will improve the analysis of NLP.

SESSION 4 Algorithmic Bias and Ethics

Moderator

Paul Han, M.D., M.A., M.P.H.
Office of the Associate Director, Behavioral Research Program, DCCPS, NCI

Patient Perspective: Kimberly Richardson

Panelists

Alex London, Ph.D.
Carnegie Mellon University

David Magnus, Ph.D.
Stanford University

Emma Pierson, Ph.D.
Jacobs Technion-Cornell Institute at Cornell Tech

Tina Hernandez-Boussard, Ph.D., M.P.H.
Stanford University School of Medicine

Dr. Alex London – Ethical Issues in AI Across the Lifecycle

Dr. London began the session with an overview of the ethical issues that can arise across the lifecycle of AI development in medicine. The term *lifecycle* references a series of key stages in the production of AI or ML systems, including problem formulation, data acquisition, model training, and algorithm validation or testing through clinical deployment. He highlighted health system features that can be problematic for AI, and how health care may need to change functionally to make better use of ML.

Initially, questions need to be posed that will ensure innovations that produce social value and advance the interests of patients, such as addressing inequities and treatment disparities. Another ethically appropriate approach would be making health care more effective or efficient by enabling stakeholders to do more with fewer resources. Unfortunately, AI tools can obscure obtaining these key goals.

For example, genetics is increasingly used as a guide to disease etiology and drug development; however, genomic databases are predominantly White and European. In 2016, only 3% of participants in genome-wide association studies were of African ancestry. AI is then built from ML systems that lack data heterogeneity, and these errors can be transferred to the tool that the researcher is creating. Clinicians cannot learn about populations who are not represented in the data, and to date, there are significant disparities in medical database diversity.

Lack of representation for various racial groups is likely to translate into knowledge and care gaps that can perpetuate or exacerbate health disparities. In one study evaluating automated methods for evaluating COVID-19 from x-ray images, researchers found that neural networks (AlexNet) trained to detect pneumonia were better able to detect where the images originated rather than the relevant pathology.

ML outputs need to ensure that meaningful goals are advanced, but this can be difficult since current AI systems excel at only classification and prediction. Clinical targets involve intervening in a manner that will benefit the patient, and predictive systems should support this goal. In another study, Obermeyer et al. examined racial bias in an algorithm used to predict healthcare cost at the time of admission with

the provision of standard medical practices. The authors found that Black patients were often sicker than their White counterparts because of inequities in healthcare access, and thus there were differences in the projected cost category. Using prediction in the domain of cost to guide interventions resulted in systematically prioritizing sicker Black patients.

ML performs relatively narrow tasks and will not produce significant social value if system outputs are not carefully aligned with evidence-based practices. Therefore, the next step is to produce the necessary evidence to establish the relative clinical merits of an AI algorithm over existing alternatives. However, although there are many ML models in development, few are validated through prospective trials or linked to meaningful clinical outcomes. This problem was illustrated by the proliferation of systems designed to code diagnostic imaging, where none were found to have clinical applicability. In another study of the use of AI to evaluate high-risk devices requiring US Food and Drug Administration approval, investigators found that there were few prospective trials, performance measures were not clearly linked to meaningful clinical outcomes, and there was a high risk of bias. Despite these issues, 9% of these studies claimed that the algorithm could be implemented in clinical practice. Consequently, AI performance claims do not necessarily align with the needs of real-world stakeholders and pose a threat to the medical goals of AI use by presenting alternatives that are not completely understood. Validation of ML systems is not sufficient to show clinical value; personnel must be trained and the system integrated into clinical workflow with performance monitored over time. Clinicians must have a plan for retraining algorithms with the changing patient populations and dynamic circumstances characterizing real-world health care. Knowledge of AI systems is still limited and can lead to misuse and proliferation of poorly validated models.

Dr. Tina Hernandez-Boussard – Building Trust Through the Intersection of Data and Technology in Healthcare

Dr. Hernandez-Boussard discussed AI applications in cancer, including diagnosis, risk prediction, treatment selection, and patient monitoring. She also reflected on methods to build trust with the use of data and technology. AI development can provide personalized medicine, but stakeholders should consider the reliability of these applications across patient populations. For example, data on heart attack patients were based on clinical trials that included only information from male symptoms, which differs from female experiences. Thus, the reliability of the data for female patients is questionable and may not be useful for detecting heart attacks in the emergency department. Similar problems arose for evaluating skin lesions in dermatology, where evidence showed algorithms performed well only for fair-skinned patients but underperformed in darker skinned patients.

As a result, there is bias in AI which systematically and unfairly discriminates against certain individuals or groups. Harms include allocation or withholding of opportunities or resources to certain groups, and representative harm when systems reinforce the subordination of some groups by identity. Bias can enter the system at various points, including data acquisition, model development, model deployment, model evaluation, and data creation.

To make algorithms more reliable across populations, researchers should consider stakeholder engagement, data quality, technology development, and transparency. Stakeholder engagement in the design and development of the model is important and includes knowledge experts, decision makers, and beneficiaries. There can be bias in model development, therefore understanding data creation, such as demographic representation, is important for laying the groundwork for ethical concerns that could emerge during the creation of AI models downstream. Data acquisition is next and involves obtaining and consolidating appropriate data for algorithm development. Model tuning includes reducing bias, such as group and equalized fairness, and overall accuracy and treatment equality:

- Group Fairness – Equal positive predictive rates across groups.
- Equalized Fairness – Both groups have an equal chance of being positive.
- Overall Accuracy Equality – Probability of prediction being correct is equal across groups.
- Treatment Equality – The ratio of false positives to false negatives is equal across groups.

Calibration helps model performance across populations, ensuring outcome applicability in the real world. Diversity is necessary to ensure that the model is fully identifying the outcome of interest. There also needs to be transparency in all phases of ML system development:

- Model Design – Stakeholder involvement, data quality, and clinical utility.
- Model Development – Output, target user, standards, model architecture, and features.
- Model Evaluation – Missing data, class imbalance, and algorithmic fairness.
- Model Reporting – MINimum Information for Medical AI Reporting (MINIMAR), code availability, and data sharing.

Often models are developed to ensure equality across groups; however, equity is more important as certain populations may require more resources to address outcomes in cancer care.

Dr. Emma Pierson – Using Machine Learning to Reduce Health Inequity

Dr. Pierson is an assistant professor in computer science in population health sciences at Cornell Medical School. She spent the past few years studying bias in ML and how this technology can worsen the issue. She presented research that she and her colleagues published in the journal ***Nature***, illustrating bias in research related to osteoarthritic pain. It is unknown why pain often persists despite diagnostic findings of disease severity. There is the possibility that the provider is missing an additional signal found in testing that could assist in reducing patient-reported pain. Traditional measures of disease severity are based on studies conducted decades ago, with data primarily from White, British populations. Therefore, it is plausible that severity measures may not capture all relevant features for different racial and ethnic groups.

In diverse populations, environmental and occupational risk factors may vary, so ML approaches can search for other features in the x-ray that might hold explanatory power for pain experienced in underserved patients. The algorithm found additional factors that were relevant for pain prediction, showing that traditional measures fail to capture all relevant signs. This information is important because clinical severity influences surgical recommendations. Hence, deep-learning algorithms can potentially create a tool to help with equity in the allocation of surgery by identifying patients with physical damage that may have been missed by traditional measures.

ML provides the unprecedented ability to expand medical knowledge, focusing on racially and socioeconomically diverse groups to train algorithms. One of the biggest lessons of this work is the need to evaluate algorithms to baseline standards and the biased world in which the models will be applied to improve clinical decision-making. A challenge is to create a model that interfaces well with clinicians in high-stakes settings with rigorous assessment of performance by all stakeholders.

Dr. David Magnus – Ethical Challenges in the Use of AI in Precision Health

Artificial intelligence is of interest in the medical community for many reasons. Physicians are bound by the number of variables they can analyze, and data are often limited to sporadic provider encounters rather than continuous data streams. There is also bias in physician practices, which points to methods that can potentially provide more targeted therapy. Building AI models that use data from multiple sources, including research repositories, secondary data uses, and consumer data, can improve predictive analytics in various domains to help guide decisions.

However, there are ethical challenges that can arise from using AI in precision medicine. Dr. Magnus discussed some of the issues he raised in a New England Journal of Medicine article he published with colleagues. There are values and interests embedded in algorithm design, biases and limitations in the data, and inappropriate reification of results. There are decisions made when building models that influence outcome optimization and weighting. Third-party vendors who sell algorithms to health systems could value different outcomes than patients or providers, including insurance status and quality metrics. Biases and limitations in data algorithms learn from missing values from underrepresented populations. These

biases can then lead to self-fulfilling predictions. Populations from African ancestry are still being omitted from data sets, which increases variations and causes problems in algorithm application to diverse groups.

Other vulnerable groups can also be at risk. For example, there is large variation across transplant programs regarding including neurodevelopmental status in transplant list decisions for pediatric patients. Some programs consider neurodevelopmental delay (NDD) as an absolute contraindication to organ allocation. Thus, at these institutions, all of these children die, even though transplant data show no difference in graft survival. Algorithms will learn that NDD is associated with death and poor outcomes, even though mortality is largely a function of biased decisions that are reinforced and reified by ML.

The differences in health outcome accuracy can be a function of societal injustice. Real differences exist in health outcomes correlated with the social determinants of health. AI can learn these traits and predict worse outcomes for minorities and other vulnerable populations, which will reinforce existing inequalities and disparities. There are better health outcome predictions for every acute and chronic disease when both socioeconomic and racial factors are considered. Algorithms can be adjusted, with some loss of the ability to predict mortality; however, the problems may not be fixable because they are tied to deeply rooted problems of social inequality in healthcare delivery.

Algorithms also can be misused because:

- Clinicians and health systems are not aware of data limitations or value decisions embedded in the algorithm.
- Algorithm results may be applied to individuals in populations where they are less predictive.

Traditionally, health care is based on the dynamic relationship between the provider and physician or healthcare system; however, now ML algorithms play a role. Patients will want to know the study group upon which clinical decisions are based. Stakeholders have a mandate to reframe relationships and obligations toward a relationship between a community or patients and providers, but our healthcare system was not established for this type of partnership.

Kimberly Richardson – Black Cancer Collaborative; Patient Perspective

Ms. Richardson stated that NCI is bringing the patient perspective directly to these kinds of discussions, and this is much appreciated by patient advocates. AI should be a tool for improving patient care, but that depends on perspective. Hospitals want quick decision-making and insurance companies want diagnostic accuracy, while patients just want to feel better. Therefore, she appreciates deliberations around prediction versus intervention and the importance to AI. Patients want to know when physician autonomy is limited, and those with low health literacy culturally rely on the physician to provide the correct answers. No true relationship exists between patients and doctors because patients cannot verify, edit, or challenge clinical notes.

Another striking example is related to gynecological cancer care. Even as recently as the 1990's nurses still thought Black women had a higher pain threshold, compared to White women. Therefore, AI could continue to collect and analyze data from a faulty model design. Justice with regard to the types and use of these tools to mitigate stereotypes (e.g., pain studies) is extremely important. AI can assist with clearing obstructions and improving healthcare inequities for vulnerable populations.

Discussion

Dr. Han asked the panelists to consider algorithmic bias in the application of ML to health care, research, and practices that may differ from conventional statistical prediction models. Dr. Hernandez-Boussard stated that in traditional statical methods, clinicians choose the variables they believe are predictive based on causal inference or clinical knowledge related to outcomes. In ML, for example, 1,000 different features may be included as variables, and the algorithm decides the importance of each. However, the

ML algorithm will only learn from the data we provide based on its representation and whether enough information on subpopulations is included.

Dr. Pierson stated that researchers know training data sets is problematic, and that fairness cannot always be achieved. ML algorithms are complex, and there is often a fundamental failure in our ability to understand the power of what we are deploying. Dr. Magnus stated that another difference is related to the social interpretation of AI compared with standard methods, which was shown with gene therapy. These tools may lead to overinterpretation and reunification of outcomes.

Dr. London stated that the computational nature of these models highlights data sources and limitations. He believes bias is an issue but that taxonomy should be broadened in terms of other problems to avoid. Bias is also becoming an exercise in applied mathematics, where the many technical definitions can be included for ensuring that health systems work equally well for diverse populations. There should be concerns around the quality and limitations surrounding the lack of diversity in data sources, which is also important from a regulatory standpoint. Problems of justice arise when diverse groups are not included in research. Dr. Hernandez-Boussard concurred with Dr. London and stated that there are opportunities at the institutional review board level and related to NCI funding for grants.

Panelists were then asked to consider how to mitigate or reduce algorithmic bias, including racial and ethnic discrimination. Dr. Hernandez-Boussard stated that her group removes race and ethnicity from their algorithms since there are many proxies for these factors, such as socioeconomic status. Dr. Magnus noted that there are other markers of race and ethnicity, but health is a marker of both. For example, type 2 diabetes mellitus is a marker of being Black and poor, and there are very few genetic diseases that are exceptions. The prevalence and outcomes of most diseases vary based on race, ethnicity, and socioeconomic status. These health disparities will be discovered by ML and may lead to differential discrimination in outcomes from predictive analytical approaches. Dr. London expressed concern about removing demographic categories because it is important to understand whose outcomes may have been influenced by medical practices. Dr. Pierson stated that algorithms can reconstruct variables such as race through proxies such as imaging and other domains. Race matters because it relates to health outcomes and links to biology are relevant. It is unknown whether including race can relieve or exacerbate health disparities.

After discussion of the limitations and hazards of the uncritical use of ML, panelists were asked to reflect on whether ML is ready for real-world application and what research might be necessary to prepare. Dr. Pierson stated that she could not evaluate the AI field as a whole in both academia and industry. There are many algorithms that are not ready for use. Clinical trials using ML are promising, and their usefulness depends on whether they are explanatory or pragmatic trials. Dr. Magnus stated that algorithms used outside the group in which it was developed may exacerbate problems of inequity. Studies start with proof of concept and take time for evaluation prior to widespread clinical use. Dr. London stated that AI is a tool that may allow the use of EMR data to create a more efficient health system, but that there is a huge gap related to system integration into clinical pathways.

Dr. Han asked what work could help stakeholders get to the point where ML can be successfully implemented, and how diverse stakeholder needs and expectations could be addressed. Dr. Magnus stated that, based on focus groups and survey research, there are many false perceptions of the basis for clinical decisions. Patients believe, regardless of racial or ethnic background, that clinicians are making an individualized decision about what is best for them, and the providers have autonomy. These misperceptions persist, despite efforts to educate to the contrary. Systemwide changes need to be made to implement ML. Dr. London stated that there is a great deal of uncertainty concerning AI medicine, and many domains which are unexplored. More work needs to be completed before ML can be trusted to make clinical judgements. Dr. Hernandez-Boussard agreed with Ms. Richardson that no real doctor-patient relationship exists. Patients do not know if the information they provide is incorporated into the EMRs, which would then be included in data factored by AI.

SESSION 5

Unstructured Data: Drawing Meaning Across Domains to Inform Healthcare

Moderators

Elizabeth Hsu, Ph.D., M.P.H.
Surveillance Informatics Branch, Surveillance Research Program, DCCPS, NCI

Nicole Senft Everson, Ph.D.
Health Communication and Informatics Research Branch, Behavioral Research Program, DCCPS, NCI

Patient Perspective: Kimberly Richardson

Panelists

Guergana Savova, Ph.D.
Boston Children's Hospital

Jessie Chin, Ph.D.
University of Illinois at Urbana-Champaign

Meliha Yetisgen, Ph.D.
University of Washington

Zhijun Yin, Ph.D., M.S.
Vanderbilt University Medical Center

Dr. Guergana Savova – An Overview of Natural Language Processing (NLP) Methods Development and Applications in Biomedicine

Dr. Savova provided a broad overview of state-of-the-art methods for NLP or text processing, and applications in biomedicine, with a focus on ML. The example she presented came from a project funded by NCI on extracting deep phenotypic information on cancer patients and merging with structured data for translational science studies. Information is extracted from health-related free text through collaborative development of innovative computational methods and software development. Information from the EMR is summarized and saved in a database, then a query is made of both structured and unstructured data.

The Apache Clinical Text Analysis and Knowledge Extraction System (cTAKES) software determines downloadable information from sub- or unstructured documents, analyzes through NLP and stores the information in a database, which is then used for cohort training. The first section covers DeepPhe, which enables document classification concepts related to identification of phenotypes. NLP uses ML models with text matching patterns and rules to identify items on cancer attributes. The user sets the values, clicks run, and output is used for cohort identification. The extracted cancer information is then organized into a cohort for identification of workflow. The workflows consist of information on patient age and agent, cancer stage, diagnostic information, and biomarker summaries with interactive features allowing for data granularity for specific cohorts. One powerful feature of this technology is the ability to identify and classify facts about a patient's cancer, such as basic information validation and cancer attributes.

The next topic concerned the temporal relationship of extraction, which is a challenging task. Dr. Savova has worked on this collaborative project since 2010 with investigators across the country. She showed an example of the complexity of the problem, as well as methodologies used to improve the granularity of

time-related text to establish links between events and dates. The EMRs have a specific document creation time so the event can be related to that time, but fine tuning of the ML domain to learn specific tasks is required. They are distributing this information through the Health Natural Language Processing Center, which is restricted to clinical data with daily use agreements, following all necessary steps for distribution.

Feature-rich methods have input with features specified or defined by the engineer or researcher, but extractors of those features must be developed. Thus, support vector machines are used to devise these features, and they also delved into neural approaches, such as CNN, encode timex, and syntactic embedding, to further improve results for temporal relation extraction. In 2018, they started experimenting with Long Short-Term Memory, another neural architecture, but now the state-of-the-art is transformers or encoders such as BERT. Domain pretraining and domain fine tuning involve training on billions of parameters or words, using data to define the probability gradients to specific tasks.

Dr. Meliha Yetisgen – Clinical Natural Language Processing for Secondary Use

Dr. Yetisgen discussed clinical information extraction and the high-performing models used to access texts on coded health information. The approach her organization uses is data-driven ML, which produces high-quality data sets. The EMR contains structured data and clinical text. This unstructured clinical text is extracted into a structured representation, which has secondary use applications such as large-scale studies and real-time clinical decision support.

Clinical text documents are created by clinicians and include patient history and treatment, containing > 80% of patient information in the EMR. This information requires interpretation or extraction. Ongoing clinical information extraction projects include patient-reported symptoms, social determinants of health, and radiology findings. Secondary use projects in the cancer domain include extraction of symptom burden from clinical narratives of cancer patients, development and pilot testing of a measurement tool for early cancer diagnosis, and large-scale clinical and economic impact analysis of potentially malignant incidental findings in radiology reports.

Clinical NLP research workflow includes de-identification of the research data set from the EMR, annotation of gold-standard corpora, developing the NLP framework, and adaptation to the domain. De-identification or sharing of annotated corpora has been challenging in the clinical domain to preserve patient confidentiality. Issues include the transferability of de-identified models across institutions, and a manual check of gold-standard corpora for possible de-identification errors.

In addition, the quality of gold-standard annotation determines the success of the NLP system. Annotation is based on the clinical task and incorporates creating a detailed set of annotation guidelines or tools, training annotators, then annotating the text. Text annotation should be based on guidelines, with the revision of those guidelines if needed. There also needs to be monitoring of inter-rater annotator agreement and re-training of annotators. Annotation should continue until the desired numbers of documents are completed. They do considerable work on the transferability of these models across institutions as performance will vary.

Event-based annotation characterizes clinical events across multiple dimensions such as status, extent, and temporality. The trigger anchors and disambiguates the events; arguments are labeled by argument type, span, and subtypes. Span-only arguments also include argument type and span, which is not easily normalized. Symptom event schema include trigger, assertion, change-of-state, severity, etiology, anatomy, characteristics, duration, and frequency, while social determinants of health event headings include employment, living status, insurance, substance abuse, and physical activity with various schema such as trigger, status, duration, history, and type included under each heading. Radiology event headings include indication, lesion finding, medical problem finding, and follow-up recommendations with associated schema.

Annotation is difficult since there are millions of notes, but limited data secondary to budgets. Random selection of notes is suboptimal since the phenomena of interest may be infrequent and many samples

may be similar. Active learning identifies samples that maximize model learning and is well established for single-label tasks. It is less established for more complex event extraction. The NLP framework analyzes the task and trains the data set. A review of existing work and defining a strong baseline is important, followed by the development of architecture and the evaluation of the test data set.

Dr. Zhijun Yin – Understanding Health-Related Behavior Using Online Data

Dr. Yin is an assistant professor of biomedical informatics at Vanderbilt University Medical Center. He stated that it is important to recognize differences between association and prediction, and between ML and statistical inference. In referencing natural language processing of text, it depends on the focus – association or prediction. There is abundant information following a clinical visit including ICD codes, Current Procedural Terminology codes, lab tests, drugs, and clinical notes; however, we know little information about patients prior to discharge.

Patient information can be obtained from surveys and interviews, online social media platforms, and clinical communications. Surveys are confined by predefined questionnaires; reduce collection response bias; and they are expensive if you want to want to conduct a large scale on survey. Online social media platforms provide large-scale data, but there is misinformation and data cannot be easily linked to EMRs. Clinical communication can be linked to clinical facts, but there is a lower adoption rate. Using online data, surveys can obtain the study cohort, while online data provide information to build features and identify outcomes.

For example, a study in caregiver needs in multiple communities related to Alzheimer's diseases and related dementias (ADRD) consisted of two hypotheses:

- Build an efficient classification system to identify online submissions on Reddit related to ADRD care challenges and experiences
- ADRD care information disclosed in the non-ADRD communities differs from the ADRD community

Tasks consisted of data collection, including cleaning the data. There was then an initial screening of the information to obtain ADRD-focused data using gold-standard references, followed by annotation and classification. The best model was chosen for ADRD care. Traditional and deep ML were used to fine-tune classification, while structural topic modeling was used for the second hypothesis.

Other research involved hormonal therapy discontinuation and clinical communications. The study cohort was built using ML to extract discontinuation (interruption versus completion), and then they investigated messaging patterns, including communication between the healthcare providers and patients, while controlling for various demographic and treatment variables. They analyzed message volume, rate, and topics using word clustering, then did survival analysis based on medication discontinuation events. They extracted posts temporally and built a social graph with reciprocity and also extracted emotion scores and other control variables. Limitations include an algorithm built only on a population from one social media platform, and limitations of statistical inference.

However, since data collection is difficult, it may be useful to look at other sources, such as online platforms, to obtain information. Researchers should rethink bias issues, information leakage, and combine different types of patient voices. A decentralized data collection model may be useful to climb the ladder of causality.

Dr. Jessie Chin – Learning from What Patients Said: Connecting NLP/ML to Patients' Language in Informal Setting

Dr. Chin has a background in psychology, cognitive science, systems design engineering, biomedical and health information sciences, and information sciences. Her research focuses on leveraging ML reprocessing to translate information in behavioral and social sciences to propose solutions in health care. She is interested in learning about what patients say about health from social media and

conversational data, which can be a proxy for patients' attitudes, knowledge, and behavior intention in a naturalistic setting.

There are two applications for NLP/ML, including health misinformation on social media. She described an example of the low uptake rates of human papillomavirus vaccines in the United States and associated misbeliefs about adverse events, side effects, and infection likelihood. They used various computation models to identify high-risk information on three domains – unverified information (identification and building the veracity continuum), underlying mechanisms (psycholinguistics, content, network, and system), and impact (size of cascade, risk perceptions, and coverage). Results showed that the convolutional neural network model – or NLP/ML model – outperformed all other models. True and false health messages rely on different mechanisms to create impact and speak differently. There were variations in semantics, such as processing fluency, negativity, topic, and vocabulary diversity, and framing of risk.

The second application was promoting health behavior with conversational AI agents. In this research, they developed a chatbot for delivering brief motivational interviewing (MI), known as MintBot. There were increasing concerns about COVID-19 vaccines for patients with compromised immune systems, such as multiple sclerosis patients. Thus, they conducted a one-time, 15-minute study using a brief MI and a simulated conversational agent with annotations. The MI consisted of open-ended questions, affirmations, reflective listening (empathy), and summarizing using NLP. Preliminary results showed that some strategies were more likely to be used in the short counseling sessions, with initial modeling showing that certain strategies, such as questions, affirmations, summarizations, and reflections, were more accurate for identifying a client's intent to be vaccinated. The LotMintBot was the development of a chatbot for delivering long-term MI, which was studied on hemodialysis patients. Similar strategies were used to develop the algorithm, but the study lasted 12 weeks.

Key points and recommended practices from this research include defining or operationalizing data structure, human annotations, patient-centered knowledge representation based on the patients' vocabulary, and patient-reported health outcomes, such as direct self-report measures and estimates through ML/NLP. Limitations from the work include incomplete/unstructured linguistic data, the need for human annotations (implicit/ambiguous meanings and hidden user intent), the need for domain-specific knowledge, and lack of direct clinical outcomes. Potential benefits from ML/NLP to inform health care include multi-level models of population health, individual and population surveillance, discoveries of socio-behavioral determinants in epidemiological models, new approaches to deliver health care and eliminate disparities, and new methods to promote dissemination and implementation of evidence-based medicine.

Kimberly Richardson – Black Cancer Collaborative; Patient Perspective

Referring to information presented on a patient with a diagnosis of breast cancer who had brain metastasis in less than 2 years, Ms. Richardson stated that she wondered about documentation of the patient discussing symptoms, such as headaches, that might be related to this medical event. This information is also relevant to expressions of time, particularly related to cancer recurrence. She also is interested in the future of precision medicine and how epigenetics may play a role in the future of temporal expressions.

Ms. Richardson stated that sometimes clinicians appear to misunderstand the power of social media or the lack of social media support groups. She stated that these groups are very savvy and can communicate misinformation. For example, prior to her ovarian cancer diagnosis, she posted symptoms on an online support group site and was given inaccurate advice. So when she eventually received the correct diagnosis, she went back on the site to communicate accurately. She also advocated for NLP since data derived from unstructured sources can have improved accuracy over structure entries in EMRs. Ms. Richardson stated that it would be interesting to see more clinical text regarding medical events as they move toward earlier diagnosis. However, she is concerned about manual annotation and the potential to introduce implicit or semantic bias. There also needs to be education regarding the social determinants of health

and patient-provider relationships, including survey designs with simple methodology. It is imperative that cancer centers talk about adverse events as well, since most cancer centers are not required to have survivorship care.

Discussion

Panelists were asked to reflect on how patients' voices and preferences should be built into NLP/ML. Dr. Yin stated that, in his social media research, he focused on people talking about their housing conditions, but would like to link together various methods, including surveys and personal accounts, or healthcare data such as EMRs, with online information. Dr. Yetisgen stated that they extract symptoms from clinical narratives, so it is the providers representation or inference of what they heard. They would like to take information from clinical narratives for the purpose of early diagnosis in high-risk patients, but also look at post-diagnosis data to understand how symptom burden is affecting patients related to treatment or cancer progression. Data can be extracted from clinical texts, as well as other narratives, like the EMRs, or with integration from social media. Dr. Chin stated that they have difficulty using this information to assist clinically but believes the best design approaches involve patients and physicians. There is always a need for annotation of the data to train the model.

Dr. Hsu also asked the experts to comment on data annotation and the introduction of bias into active learning models. The data are constantly changing and they come from EMRs or online communities where natural conversation is evolving. So how are models updated and how can data be annotated without gold-standard references? Dr. Savova stated that everyone does manual annotation work, and even with the best AI models, gold-standard labels are necessary. He stated they are careful as to who they hire to do annotation, and they use domain experts involved with young patients to develop their annotation guidelines, applying methods for internal validity. They perform abstract meaning representations, which are high-level semantic representations requiring the domain expertise needed linguistically. With clinical data, they have in-house experts who do the foundation work. For active learning, there are many caveats because model training must be based on the predictions of that model alone. The state-of-the-art is not farming methods but using labeled data. Thus, much research needs to occur, and neural networks are being studied to edit errors found after training a very large model.

Dr. Yetisgen replied that the gold standard must truly represent the nature of the data without bias to accurately measure system performance. Much can be accomplished by leveraging the unlabeled data into a language model, but some manual orientation may need to be completed. In the end, training of the set determines the output, and the model will learn from that information, therefore researchers must be careful with active learning. Dr. Yin stated that the first principle of ML measurement is good data plus simple algorithms lead to better coding and high-quality designs. Modeling of NLP is a challenge because there are many different ways to speak on the same topic. Dr. Chin agreed that there are many variances in human language, especially in how people describe their concerns, thus both experts and patients are needed for this work.

Lastly, panelists commented on future opportunities in the field of ML/NLP and the use of unstructured data. Dr. Savova stated that even with 99% accuracy, there may still be people harmed from an AI system, and are we prepared to accept that outcome? Attention must be paid to point-of-care decisions provided by ML. Dr. Yetisgen stated that AI is an exciting but difficult area because of methodological issues, and application at the beside is still limited. Building on other researchers and providing explanations about predictions may make this information useful clinical care. Dr. Yin stated that we need different kinds of data to describe various health conditions as information is lacking in certain areas. We also need high-quality data to improve health modeling. Secondly, ML should be viewed as a new software or engineering tool requiring user sophistication. Dr. Chin stated that she focuses on how to empower patients and is interested in the combination of this research with social science.

SESSION 6

Machine Learning-Based Decision Support at Point of Care

Moderator

Bryan Kim, Ph.D.
Outcomes Research Branch, Healthcare Delivery Research Program, DCCPS, NCI

Patient Perspective: Kimberly Richardson

Panelists

Edwin Lomotan, M.D.
Agency for Healthcare Research and Quality

Kensaku Kawamoto, M.D., Ph.D., M.H.S.
University of Utah

Raj Ratwani, Ph.D.
MedStar Health Research Institute

Ravi Parikh, M.D., M.P.P.
University of Pennsylvania

Dr. Ravi Parikh – Clinician-Algorithm Collaborations to Improve Serious Illness Communication in Cancer Care

Dr. Parikh is a medical oncologist and director of the National Center for Human Factors in Healthcare. He presented on an ML project based on clinical decision support in cancer care delivery, which might serve as a model for the integration of AI and clinical care delivery. Serious illness conversation (SIC) has been formalized through various checklists and conversation guides, and sharing this information is important because it improves hospice utilization while reducing aggressive end-of-life care and spending. However, following implementation of a similar guide at his facility, only 10% of patients had a documented conversation, revealing a huge gap. Solutions that primarily rely on education fail because of patient identification and clinician behavior change.

Both topics are relevant for an integrated ML-based approach merged with behavioral economics to improve this communication. The objective of his research was to create and implement a personalized prognostic tool for oncology outpatients to increase the number of patients with a documented SIC. The ML-based algorithm was based on 150 structured EMR variables that generated real-time predictions immediately prior to a clinical encounter. Metrics included an intervention that was acceptable to clinicians, more high-risk oncology patients having an SIC, and end-of-life outcomes.

His team started with qualitative interviews to identify the problem, moving on to developing and validating an algorithm to predict the identification of patients who may benefit from these conversations. They used clinician surveys to assess face validity, with the theme of the utility of algorithms to validate intuition about prognosis, and prompt conversations about patient goals and preferences.

Dr. Parikh and his colleagues then ran a feasibility pilot study subsequent to a randomized trial. Next, they performed clinician interviews and behavioral phenotyping to identify areas of improvement. The clinicians identified concerns over algorithm accuracy, overreliance on algorithm predictions, and ethical implications around the disclosure of algorithm prediction. Prospective validation of the algorithm at

two hospitals in different cancer types showed various areas under the receiver operating characteristic curve (AUCs), with the highest in breast cancer (AUC = .93) and the lowest in neurologic cancer (AUC = .71). Clinicians also indicated that 65% of patients on their high-risk list were appropriate for an SIC that week.

The randomized controlled trial consisted of weekly SIC emails with a performance report and peer comparisons, a review of high-risk lists or opt-out patients, and automated text reminders with the default being on the morning of clinic. Findings showed fewer patients in the intervention receiving chemotherapy within the last 14 days versus the control (7.5% versus 10.4%), although hospice use prior to death and use of the intensive care unit during the past 30 days were similar between groups. Post-intervention clinician interviews from 25 oncology clinicians showed the intervention prompted better documentation of SIC in EMRs, and validated norms around early conversations. Barriers included cancer-specific heterogeneity in algorithm performance and the frequency and tone of automated text messages.

The tool appears to serve more as a reminder that it is time to have these patient conversations. However, the next steps are to improve the intervention using patient-generated data. Other active human-machine collaboration projects include precision colorectal cancer screening navigation, default palliative care referrals for high-risk medical oncology patients, and EMR reminders to improve naloxone prescribing in high-risk primary care patients. ML predictions can improve oncology care when clinicians' perspectives are integrated prior to algorithm development with vetting prior to scale. The algorithm must also be well-integrated into clinical workflows. These models support, but do not replace, decisions that are typically reserved for complex conversations. Algorithms paired with behavioral nudges rather than simply displayed on a computer screen may be better. Lastly, the phenotype of non-response can be detected and acted upon.

Dr. Kensaku Kawamoto – Leveraging Machine Learning in Interoperable Decision Support: The University of Utah Reimagine EHR Experience

Dr. Kawamoto discussed opportunities to incorporate technology opportunities into predictive modeling, such as improving EHRs and the integration of data-driven insights. Traditional EHR optimization includes configuration, training, and personalization of the system. Add-on approaches provide additional tools, similar to adding apps to a smartphone. The University of Utah ReImagine EHR initiative was a multi-stakeholder project started in 2016 with the goal of improving patient care and provider experience through interoperable EHR apps that convert actionable insight into improved care.

The system, including more than 20 solutions, was developed with $35 million in grants and won multiple national awards. Apps were created for multiple diseases and conditions, including newborn bilirubin levels, diabetes, COPD, and chronic disease management in ambulatory care. Prediction model-driven lung cancer screening was a shared decision-making app developed to determine the benefits versus harms of low-dose chest CT screening. Studies suggest that this type of screening could save more lives than breast cancer screening (10,000/year). The algorithm was driven by a 23-variable NCI predictive model of individual outcomes and improved the odds of screening referral in primary care by five times.

A chatbot was also created for genetic testing. This NCI-funded trial of more than 3,300 patients at the University of Utah and New York University, enabled by NCI Informatics Technology for Cancer Research-funded infrastructure, created a patient portal with a chatbot link. EHR outreach was included to show an encounter, along with the chatbot outcome and transcript. A testing kit was mailed based on the patient's desire for testing.

Standards-based EHR add-on apps are a powerful approach to integrating ML-supported clinical decision support into EHRs and clinical workflows.

Dr. Raj Ratwani – Machine Learning at the Point of Care: The Need for a User-Centered Approach

Dr. Ratwani stated that his formal training is in human factors and cognitive science. He spent several years in the defense industry prior to health care in 2012. Dr. Ratwani discussed the ML/AI reality gap and how these algorithms are typically developed in very controlled settings by a team of smart scientists who may not fully understand the use of these systems and how they will inform care. The designer may imagine they have a physician user, paying complete attention to algorithm alerts, when in reality providers are very busy, and constantly bombarded by competing alerts and interruptions. In emergency medicine, clinicians may be interrupted 10 to 15 times per hour or every 6 minutes.

For example, an EHR sepsis alert was created, but the provider may wonder why the model thinks the patient has sepsis or reflect on model accuracy and whether they should trust the algorithm. Challenges include explanations, as ML algorithms often do not expose the reason for a prediction. Uncertainty is also a problem, and this is related to distribution shift – where the nature of the data encountered when the algorithm is used in the real-world is different than the training data used for model development. Lastly, clinicians may not trust the model predictions, or conversely, they instill too much trust in an algorithm. Insight should be provided on how often the model is right and how often users follow the alert.

Dr. Ratwani presented safety concerns related to AI modeling. An algorithm was developed that predicted patients with asthma and pneumonia are less likely to die than patients with asthma alone. The potential outcome was less treatment for patients with both diseases, and fewer potentially significant adverse outcomes. However, the data were influenced by a policy in place which dictated patients with asthma and pneumonia would be aggressively treated and admitted, which most likely influenced mortality rates.

A skin cancer screening optimization project was also reviewed, which was comprised of a network model created to differentiate benign from malignant tumors using hundreds of thousands of images and tested against 21 dermatologists. AI technology should be developed with the end user involved in all stages of design, implementation, and long-term use. Models must be continually monitored and validated to maintain the desired performance and detect safety concerns. Models can also be retrained and re-evaluated for modification of human factor elements.

Dr. Edwin Lomotan – AHRQ Clinical Decision Support and Machine Learning

Dr. Lomotan provided a program-level overview of projects at the Agency for Healthcare Research and Quality (AHRQ) regarding clinical decision support (CDS) and ML. The mission of AHRQ is to *produce evidence to make health care safer, high quality, more accessible, equitable, and affordable, and to work within the US Department of Health and Human Services and with other partners to make sure the evidence is understood and used.*

AHRQ clinical decision support in legislation was a 2016 initiative based on patient-centered outcomes research and Affordable Care Act legislative requirements. AHRQ CDS initiatives to advance evidence into practice and create sharable standards include engaging a stakeholder community, creating a prototype infrastructure for sharing and developing CDS, advancing CDS through grant-funded research, and evaluating the overall initiative. The five "rights" of CDS include the right information, people, formats, channels, and times during the workflow. CDS should not be just an alert or reminder, or only for physicians at the point of care. CDS is a process and set of tools for improved quality of care, enabled by technology such as ML, and mostly "human."

Patient-centered CDS (PCCDS) involves prioritizing evidence for the dissemination of this information, authorizing PCCDS interventions, implementing PCCDS, and measuring PCCDS impacts. For example, MammoScreen™ provides individual guidance for patients facing decisions about mammography screening and genetic counseling. It focuses on the integration of CDS tools in EHRs and engaging patients in shared decision-making and the use of standards. There are also algorithms that can impact racial and ethnic disparities in health care. Accelerating Change and Transformation in Organizations and

Networks (ACTION) 4 is creating a digital healthcare equity framework with an accompany guide for use, slated for completion and broad dissemination in fall 2023. CDS Connect is a website that includes CDS tools and a platform to share CDS information.

The infrastructure to support sharing and learning in the CDS community includes sources such as clinical practice guidelines and peer-reviewed articles, that can assist with building CDS artifacts. This tool can be shared in CDS Connect and the artifact improved through community use. The goal of the Clinical Decision Support Innovation Collaborative is "to engage a wide stakeholder community … on how best to disseminate evidence into practice through patient-centered CDS." This work builds on our experience with the PCCDS learning network and will help identify opportunities while implementing projects and dissemination activities.

Kim Richardson – Black Cancer Collaborative; Patient Perspective

Ms. Richardson commented that patient integration is key, but personal risk factor questions must be easy to understand, secondary to low health literacy. Patients may answer questions based on what they think the clinician wants to hear as opposed to a clear understanding of the question. She also spoke about committee and expert panel creation and wondered who the experts were and who they represent. Researchers need to make sure there is representation including Blacks, Asians, and Latinos. Apps can reduce administrative burdens, but the clinicians' workflow must be considered. There are also issues when a model does not fit real-world data. How can trust be built related to this data, and will patients feel safe with an algorithm? If predictors are wrong, how will that be communicated to patients? Algorithms should not cause unnecessary harm to patients and families. Ms. Richardson also questions whether ML-generated reminders will change clinician behavior. End-of-life prediction accuracy is paramount and it can be difficult to quantify improvements on the documentation of these conversations to benefit patients and create best practices. Patients must partner on the design of questionnaires.

Discussion

Panelists discussed the current challenges of ML specifically related to cancer care, and what is needed to develop and promote AI-based learning and support at the point of care. Dr. Kawamoto replied that model generation to identify the right patients and developing an appropriate model are important to avert increased disparities in health care. Dr. Lomatan stated there are many opportunities for ML at the bedside but finding the right system at the right time is key. He stated that they are partnering with patient advocates and including diverse participants in meaningful positions. Dr. Parikh remarked that it may be difficult to get cancer care providers to use algorithmic-based decision support in routine care.

Dr. Ratwani commented on next steps when a model does not adjust to real-world data and how to build trust in the algorithm. Accuracy is a burden on the provider, and can have significant implications for patients, such as wrong treatment and adverse events and outcomes. ML systems will be wrong at times, so the issue is how to represent this uncertainty to providers. Providers should be equipped with the correct information in the best format to facilitate their reasoning and decision-making.

Regarding methods to change clinician behavior related to SIC, is there any competition among providers generated by period comparisons? Dr. Parikh stated that while it was a controversial aspect of the CDS intervention, post hoc interviews showed it was a powerful motivator toward increasing these conversations.

Dr. Kawamoto commented on the need to respect clinician workflow when designing ML algorithms. He stated that changing behavior is very difficult when clinicians are overburdened, and it is important to make the system as efficient as possible. Engaging the patient directly may be one solution; however, US Food and Drug Administration regulations state that anything done directly with patients now becomes a medical device, which may not be supportable by research or federal funding.

ML outputs are combinations of data and models, yet these algorithms are in a constant state of decay. Panelists were asked to comment on the responsibility of those building these tools to monitor and continue to update their models. Dr. Parikh stated that there is a huge responsibility in the industry to address this decay. Sometimes the drift occurs too quickly to reconfigure the algorithm; however, it is important to monitor for the sources of these errors over time and to identify inputs that may be prone to drift. Is a shared responsibility of the provider organization and adopter to work with industry to solve the problem. Many organizations do not have the resources to fix these issues and are unaware of the sustained cost of maintaining this technology. There is a reinvestment that is required to ensure usability and safety, as well as the inclusion of new features.

Dr. Kawamoto stated that the biggest benefit of ML is greater accuracy than that for traditional approaches or simple models. Dr. Ratwani concurred and stated that the opportunity of ML overrule-based approaches is related to prediction and the kind of accuracy not possible before. The biggest disadvantage is the inability to explain what is driving the model's prediction. Dr. Parikh added that other issues involve training and integrating some ML-based tools.

Lastly, panelists addressed from a research perspective how to build an evidence base for clinical care adoption of ML. Dr. Kawamoto stated that there needs to be collaboration with implementation scientists. Dr. Ratwani agreed and stated that there needs to be an understanding of the elements that a framework can hold. He also stated that a strong governance process must be included with CDS and other health information technology, along with rigorous implementation science evaluation and academia-industry partnerships.