BMIR
Stanford Center for
Biomedical Informatics Research

Connection

Fall 2021

We Connect Data to Health

What's Inside

Keeping Our Focus on COVID-19

With emergency use authorization of COVID-19 vaccines last December, it was expected that by now the pandemic would no longer be front page news. Instead, the world remains focused on research related to messenger RNA, and BMIR is part of that.

On page 3 of this issue, you can read about some exciting research being performed by the Khatri lab. Their work contributed greatly to a recent study, published in Nature, that gave powerful support to the need for the current two-dose sequence of the Pfizer-BioNTech COVID-19 vaccine. The study was based on an interdisciplinary systems immunology approach to identify a global map of complex interactions between cell–cell, cell–environment, protein–protein, and protein–DNA interactions.

The work, which required the Khatri lab to cross analyze vast amounts of collected data, is the epitome of what we do in BMIR.

Mark Musen, MD, PhD
Director, Stanford Center for Biomedical Informatics Research

ARTIFICIAL INTELLIGENCE

BMIR Physician Data Scientist Is Developing Recommender Engines to Facilitate Medical Decision Making

After Jonathan Chen, MD, PhD, completed his medical training he realized the great variation in how patients are diagnosed and treated.

"A century ago, a single doctor could know it all and do it all because we didn't really know how to treat that much. Since then, exponential advances in medical science have enabled us to take better care of people, but it has also led to tens of thousands of diagnoses, tests, and drugs. We are well past the point where the complexity of modern medicine exceeds the capacity of the unaided expert mind," said the BMIR physician-scientist.

Because of that, Chen's current research direction is focused on trying to discover and distribute the latent knowledge and clinical data through informatics and computational tools.

"Medicine turns out to be more complex than you might expect. It has to do with the scarcest resource, access to professional expertise through a clinician's time. I think that's where informatics has a huge role to play. It's really the only practical solution to the escalating overall complexity of medicine, and the only way to effectively deal

with the iron triangle of health care cost, quality, and access," he added.

Billions of people worldwide need access to medical expertise, not just in underdeveloped areas. Patients can have their care compromised while waiting months for an expert medical consultation, even in well-populated settings in industrialized countries.

"Not only that, but health care expenditures in the U.S. are more than double most other developed countries, and it's only growing as a share of our economy," Chen said.

His mission is to change that by developing fully digital consultation systems.

One technique he is working on is recommender algorithms for medical decision making, which is based on technology that Amazon, Netflix, and other e-commerce and streaming services use to recommend purchases, books, or movies based on what "others like you" have enjoyed.

In Chen's vision, physicians will use recommender engines that comb real-world clinical data to help them make key decisions based on steps other clinicians have taken with similar patients, empowering individuals with the collective experience of the many.

Chen is currently using his artificial intelligence algorithms in collaboration with Stanford's Departments of Medicine and Pathology, and he expects to expand that work into value-based care initiatives, cost-saving reinvestment programs, and other areas.

In the near future he expects decisions based on recommender engines to be commonplace in exam rooms, helping physicians become better at what they already do and making medical practice a more consistent, universal experience for everyone.

CLINICAL APPS

QSU Investigator Summer Han Collaborates with Clinicians to Attack Second Primary Lung Cancer

Although her particular expertise is in data science, Summer Han, PhD, is motivated to work with investigators who are making an impact in clinical practice.

That became evident in 2015, when Han began a key collaboration with Oncology Division Chief Heather Wakelee, MD. In Wakelee's lung cancer clinic were survivors who were developing a second primary lung cancer (SPLC). Because it was not clear who was at high risk for developing SPLC, Wakelee wanted help in developing strategies for advising long-term lung cancer survivors.

Han, a principal investigator in the Quantitative Sciences Unit (QSU), worked with Wakelee to develop a pioneering SPLC prediction model based on factors such as age, sex, race, treatment, histology, stage, and extent of disease. They evaluated the clinical utility of the model by calculating its net benefit in varied risk thresholds for screening and in 2017 published the results of a study in the *Journal of Clinical Oncology*.

The authors noted that the model did not include other key information on genetics, smoking, and other factors that could contribute to the prediction model for SPLC. That led Han to seek, and receive, a National Cancer Institute R37 MERIT Award for an Early Stage Investigator. The five-year grant is supporting research to identify the genetic, clinical, and environmental determinants for SPLC, to assess an individual's risk of developing SPLC, and to evaluate efficient lung screening strategies for SPLC to help inform the development of consensus screening guidelines for lung cancer survivors.

As a primary investigator running her own lab, she relies on the multidisciplinary strengths of biostatisticians, epidemiologists, and medical doctors on her team as she pursues studies with practical ramifications. As an example, Han had developed a mathematical model that she wanted to convert into a web-based risk assessment tool to aid clinical decision making for lung

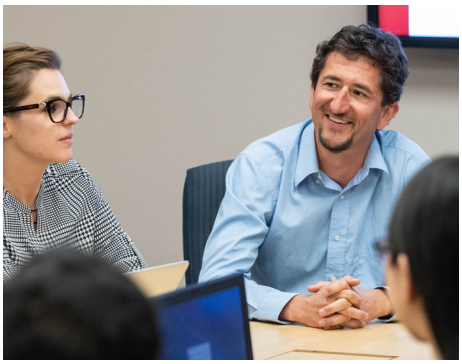
cancer patients and survivors. Results to date led to the publication of Development and Validation of a Risk Prediction Model for Second Primary Lung Cancer in the July 13, 2021 issue of the *Journal of the National Cancer Institute*.

In practice, Wakelee and other oncologists at Stanford Health Care plan to use this app in counseling prospective patients.

"I don't want the results of my work to just sit in the literature. I want to be involved in activities that have an impact on clinical practice at Stanford and beyond," she said.

CANCER RESEARCH

Machine Learning Technique Leverages Unrelated Data to Study Rare Cancers



For scientists studying rare diseases, it is usually difficult to build good predictive models because limited sample sizes and highly selective features result in a shortage of data.

Olivier Gevaert, PhD, has shown that a technique used in machine learning can be used to overcome a shortage of data in rare cancers by relying on data from more common cancers.

Gevaert was among five authors of a December 2020 article in *Nature Communications* that demonstrated how a meta-learning framework can be used effectively to leverage data that is relevant but not directly related to the problem of interest.

"We showed in the paper that we need an order of magnitude less data to reach a viable model in cancer prognosis. We can reach the same performance in the prediction of the task in the target domain with data from 20 patients compared to data from hundreds of patients using a technique called meta-learning," he said.

Meta-learning is a more advanced way of doing transfer learning. In transfer learning, researchers try to borrow data from a larger source domain that is closely related to the much smaller target domain.

When studying rare diseases like brain tumors, for example, it is often difficult to collect a significant amount of data because of a limited number of cases or because the technology to generate the data is too expensive.

Gevaert explains that "we can build models based on data obtained from other cancers like breast, lung, or prostate and transfer those models to the specific brain tumor we're studying."

Gevaert wrote more about this in a March 29, 2021 commentary in the *British Journal of Cancer*.

"Meta-learning is showing promise in recent genomic studies in oncology. Meta-learning can facilitate transfer learning and reduce the amount of data that is needed in a target domain by transferring knowledge from abundant genomic data in different source domains enabling the use of AI in data scarce scenarios," he said.

Now that the technology has been shown to work for genomic data, Gevaert wants to extend it to images because of a limited amount of data available for certain imaging data modalities.

"We have 2-D images in digital pathology and we have 3-D images in radiology for cancer patients. I want to see if we can use the same mathematical framework and show that it also works in these areas," he said.

REPRODUCTIVE HEALTH Leveraging EHR Data for Insight into Pregnancy and Birth

Research scientist Alison Callahan, PhD, is building one of the very few longitudinal datasets for improving our understanding of the health of obstetric patients.

She is using her informatics skills to analyze electronic health records (EHR) and build a pregnancy and birth patient database known as the PRregnancy Outcome Research Database (PROGRESS).

Past research in this space has leveraged insurance claims and other administrative data, but that is sorely lacking in maternal and fetal medicine information. One of the research areas motivating the development

of PROGRESS is focused on pregnancy loss, for which relevant information is very difficult to extract from administrative and claims data. She is collaborating with Stanford reproductive medicine specialist Gaya Murugappan, MD, and epidemiologist Stephanie Leonard, PhD, to profile pregnancy loss from electronic health records data at Stanford. They hope to shed more light on that subject by extracting information from de-identified EHR data such as gestational age, number of prior pregnancies and births, health history, and post-pregnancy outcomes.

Callahan is also working with Murugappan and Leonard to identify a cohort of patients who are at low risk for post-birth complications. There is great value in such a cohort for a wide variety of research projects. Examples of areas to study are rates of Caesarean section and many types of post-birth complications and their rates. "That cohort of patients is surprisingly challenging to identify in administrative data, but we are developing a method that would let you identify that cohort just from electronic health records data here at Stanford," Callahan said.

Furthermore, by using the STanford Research data Repository (STARR-OMOP), they intend to build robust methods that can be generalized for use by other health systems and members of the Observational Health Data Sciences and Informatics (OHDSI) research community, making it possible to identify much larger study populations for future research.

"By developing these resources and these methods we hope to foster research based on new questions we should be asking to improve the health and health care for pregnant and birthing people," Callahan said.

On an even broader scale, the work has implications for the more complex interaction of health care and health systems.

"A larger database will allow us to address disparities in access to care for pregnant and birthing people and the corresponding disparities in the rates of mortality and morbidity in different groups here in the United States," she said.

COVID-19 Analysis of Extensive Data Gives Insights into Immune Responses Induced by Pfizer Vaccine

The lab of Purvesh Khatri, PhD, collaborated on the first systems analysis of an mRNA vaccine and shed light on the powerful boost gained from the second dose of the Pfizer-BioNTech COVID-19 vaccine.

Khatri and colleagues described a study that was published in the July 12 issue of *Nature*.

The Pfizer mRNA vaccine belongs to a new class of vaccines, and the research team wanted to learn what molecular changes occurred in the body with the vaccine. The study was an outgrowth of their previous work, published in *Science*, that the researchers performed to model infection with COVID-19, using the same technology and the same analysis methods, known as a systems immunology approach.

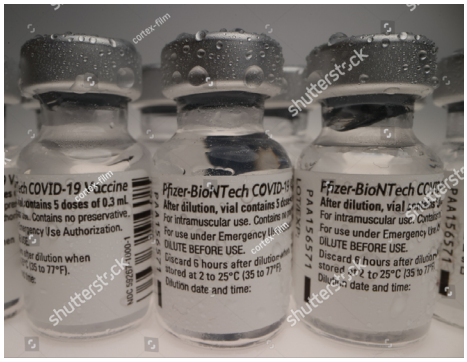
Systems immunology is an interdisciplinary approach that uses high-throughput technologies and computational methods that can be applied to identify a global map of complex interactions between cell–cell, cell–environment, protein–protein, and protein–DNA interactions.

Following analysis of a large amount of data generated, the researchers observed few inflammatory changes after a first dose of the vaccine. But after a second dose, there were 10 to 100-fold increases in inflammatory responses in specific immune cell types.

"That underscores the importance of not skipping the second dose, because all the changes that we observed were coming after just 24 hours of receiving the second dose," Khatri said.

"We also observed a very profound biological insight—that a specific subset of innate immune cells, called monocytes, have memory too," he said.

A subset of monocytes constituted only 0.01% of all circulating blood cells after the first dose of the Pfizer vaccine. But after the second dose of the Pfizer vaccine, their



numbers expanded 100-fold to account for a full 1% of all blood cells. In addition, their disposition became less inflammatory and more intensely antiviral, which indicates their ability to protect broadly against diverse viral infections.

The research also showed no evidence of auto antibody generation in any of the subjects, which suggests that the vaccine does not lead to the immune response attacking the body itself.

COLLABORATION

OntoPortal Alliance Enables Use of BioPortal Technology in Many Application Areas

BMIR software that created the BioPortal ontology repository will be repurposed to enable the creation of an entire network of ontology repositories in various application areas through an OntoPortal Alliance.

An ontology is an organized collection of concepts—a vocabulary of a very specialized sort—to publish and exchange knowledge and understanding. By following shared specifications, ontologies organize terminologies in many different disciplines, enable data to be annotated in standardized ways, and provide knowledge that allows software researchers and systems to perform natural language processing, data and knowledge integration, and information management. At their core, ontologies provide a formal means to define the relevant entities in an application area and to give those entities standardized names so that researchers can talk to each other.

Since it was created more than 15 years ago, the software that BMIR wrote to manage a repository of ontologies has gained popularity around the world among researchers using it not only for its original biomedicine focus, but also in many other areas of science. The BioPortal software is now freely available as OntoPortal, reflecting its generic ability to represent any ontology.

A consortium of researchers in the U.S., China, France, and Italy have joined the OntoPortal Alliance, which is using and advancing the BioPortal software. These researchers provide ontology services for biomedicine, agriculture, and environmental research. In addition, a group in Germany is working on a material science portal, and at least three companies have expressed interest in joining the OntoPortal Alliance as they engage in commercial pursuits.

As an example of how OntoPortal is being used, the French semantic services team, led by Clement Jonquet, PhD, of the Laboratory of Informatics, Robotics, and Microelectronics of Montpellier, developed AgroPortal, a vocabulary and ontology repository for agronomy and related domains. AgroPortal resources currently support more than 50 projects from around the world, such as BigDataGrapes, which is focused on European companies active in two key industries powered by grapevines: wine and natural cosmetics.

BMIR plays a crucial role in maintaining the BioPortal and OntoPortal software. BioPortal enables functions in mission-critical research systems such as REDCap, CEDAR, and the routine work of hundreds of businesses including Philips, Dupont, Allscripts, and GSK. The OntoPortal deployments by other users of the software provide similar services for their domain communities.

A grant from the National Institutes of Health will support BMIR's efforts to facilitate how the various repositories exchange information. The goal is to enable users to pose queries that will integrate ontology information from all the OntoPortal instances in the network.

"We really want to make this community as well established as we can," said John Graybeal, BMIR's BioPortal Project Manager.

"We have software that is standardized and manages terms in a way that people and computers can understand them. And so, the OntoPortal Alliance is about sharing our way of doing that—of providing services around the terms and how they're managed—so that everyone can be accessing common software, common resources, and common capabilities in similar ways," he said.

BMIR
Stanford Center for Biomedical Informatics Research

WE CONNECT DATA TO HEALTH

The Stanford Center for Biomedical Informatics Research (BMIR) uses advanced research techniques to discover, apply, translate, and organize data that make a difference for health and health care. With its expertise in clinical and translational informatics research and biostatistics, the division works to uncover new ways to advance personalized medicine and to enhance human health and wellness.

Collaboration is in our DNA. We are excited about the prospect of working with other experts who share our goal to connect data to health and medicine. We encourage you to contact Mark Musen, Director of BMIR (musen@stanford.edu), to learn more.

Connection
Fall 2021

Stanford Center for Biomedical Informatics Research
1265 Welch Road, Stanford, California 94305-5479
https://bmir.stanford.edu