Cancer Challenge Area 1

Generating Synthetic Data Sets

Every year in the United States, petabytes of clinical oncology data—from PET scans to biomarkers—is collected from tens of thousands of patients in individual hospitals and research centers[i]. However, due to patient privacy and other data sharing issues, most of this data is not publicly available or shared and is thus unable to be leveraged by the broader research community. Meeting participants identified this "data jail" dynamic as a "foundational" hurdle to advancing computational oncology.

Computer-generated synthetic data sets are statistically identical to real clinical data sets but are anonymized and are thus not considered protected data. The vision at the heart of this lean-in challenge area is an ecosystem in which original "gold standard" data sets remain under stewardship of the entities that create them, such as cancer registries. The synthetic data set versions created from these gold standard data sets, together with their specific generation rules and metadata, would be made broadly available. These distributable data sets would provide researchers with the ability to apply new models of analysis to these synthetic sets, and subsequently offer the products of these analyses, such as new machine learning (ML)-algorithms, back to the original owners of the clinical data sets and the broader research community.

"If there is the ability to develop synthetic data sets, that allows the original data set to remain incumbent or embargoed where it stands, such as your clinical trial data, then that derivative data set can traverse the world along with its methods, metadata and labeling and that allows users of that synthetic data set to use it for whatever they wish."

-- Nick Anderson, Ph.D., UC Davis

Relevance

Current clinical data sets that are shared are typically of small scale, have significant data attestation rules and requirements, and reflect local inconsistencies in data from source systems.

Access to and sharing of sufficiently large and high-quality comprehensive clinical data sets is inefficient and difficult to incentivize. Gaining access to and managing clinically derived data is time-consuming, expensive, and de-identification is subject to multiple intellectual property (IP) and institutional review board (IRB) restrictions. In addition, de-identification of these data sets varies between institutions. Even when access is achieved, data quality of extracted clinical data sets often suffers due to "over sanitization" in the de-identification process, and inconsistent labeling. This makes it challenging to generalize results to further modeling, training or evaluation needs.

Synthetic data methods and resulting data sets promise to protect patient confidentiality by completely delinking identity. Other benefits include advancing the expectation of transparent and reproducible analysis methods and supporting transferrable methods that can be applied to original data for clinical validity. An ecosystem that supports the access to synthetic data sets could both expand the scale of opportunity for education and training on complex clinical data and deepen our understanding of the complexity in healthcare data access and management.

Why Now?
Meeting participants noted that computational technologies such as ML, and changing community standards (open-access and data sharing policies) combined with pioneering efforts in clinical synthetic data make this an ideal time for a larger oncology-based program.

For example, MIMIC-III is a large, publicly-available database comprising de-identified health-related data associated with approximately sixty thousand admissions of patients who stayed in critical care units of the Beth Israel Deaconess Medical Center between 2001 and 2012[ii]. Similarly, there has been a precedent set with the large-scale sharing of cancer genomics data via the NCI Genomic Data Commons[iii]. These two large-scale data sharing projects are driving expectations for the greater sharing of clinical oncology data.

Innovation Potential for Cancer Research
The ability to produce anonymized or de-linked synthetic data could drive the capture of a wider range of clinical data not currently recorded. The sharing of synthetic data, their generation methods and metadata could be critical to improving the reproducibility of clinical data since different synthetic data sets methods could be tested against one another using methods developed in other fields and subsequently inform improvement in clinical data quality in a learning healthcare system.

This ecosystem would also advance the ability to develop and share complex computable phenotypes of patients among medical environments and establish new ways to determine missing data or noise though peer reproducibility and validation. In addition, it would contribute to improvement and standardization of quality measures for source data systems and clinical workflows with provenance provided back to gold-standard data generators.

The access and validation of new or different ML or other analytic models on accessible clinical datasets of any size would advance democratization of data and model transfer. It would also inform how clinically-relevant gold standards of data sets support higher standards of data reproducibility.

Innovation Potential for High-Performance Computing Capabilities
Providing an ecosystem of synthetic data sets based on real, clinically-associated data will require new, innovative collaborations in ML, clinical data generators, data managers, data architecture and software engineering.

Anonymizing large data sets (100,000s of patients) and ensuring these cannot be used for re-identification of individual patients will require new advanced algorithms and an expanded expectation for stewardship of the full workflow process involved in large-scale clinical data-set generation.

Cancer Challenge Area 2

Machine Learning for Hypothesis Generation and Clinical Decision Support

"What is needed are computational models that can generate a hypothesis from the existing data that then can feed back into the experimental loop making cancer research and clinical trials (e.g. adaptive clinical trials) more efficient."

-- Jeremy Goecks, Ph.D., Oregon Health & Science University

Meeting participants identified machine learning (ML) as a critical tool for experimental hypothesis generation and guiding experimental design. ML, a type of AI, provides computer systems the ability to automatically learn and improve from experience without being explicitly programmed, and ML algorithms build a mathematical model of sample data to make predictions. Thus, ML is a cornerstone computational tool for the movement from descriptive to predictive approaches and eventually to realize the potential of precision oncology—the right treatment, for the right patient, at the right time.

Relevance

Meeting participants noted that at present there is a cancer data conundrum: researchers and clinicians are inundated with more information than they can handle, while, at the same time, there are sizeable gaps in the biological systems information that is required for research and clinical advance. ML is key to addressing both hurdles. First, ML can provide guidance to experimentalists on what to study and the sequence in which to attack questions. Thus, ML models have the potential to guide experiments and fill out data gaps as efficiently as possible. Second, ML is a critical bridge between large, complex data sets and mining actionable meaning from the data. ML has the unique ability to discover and use algorithms to cluster observations (data), and to do so iteratively with experimentation, in an active-learning process. As such, ML has high potential to be used to develop tools to support real-time clinical decision making.

Why Now?

In the past decade there has been a fundamental shift from qualitative to quantitative data in cancer research and clinical settings. This ongoing analog to digital transition is producing an extensive pipeline of previously unavailable digital data related to cancer. For example, scanning electron micrograph (SEM) technology has transitioned from photographic data ("*blobology*") to fully digital, near atomic-level resolution scans.[iv] As such, data has leapfrogged application.

Already, exciting new research has demonstrated that ML and other computational tools can be used to analyze digital histological data—and critically, to do so with greater diagnostic accuracy than clinicians alone[v]. Early results from the JDACS4C collaborative cancer research pilots show promise in applying ML for hypothesis generation. Likewise, advances in Deep Learning, a branch of ML based on artificial neural networks, make this an opportune time to apply ML on leadership-class HPC platforms for large-scale analysis of public data sets. DOE's new Summit leadership-class supercomputer has been designed for ML and AI applications and, as of January 2019, is available for cancer research applications.

Innovation Potential for Cancer Research

Meeting participants noted that applying ML would advance cancer research and clinical applications in two key ways. First, it can be used to generate novel hypotheses to determine what additional experiments/data are needed to improve clinical outcomes, understand cancer systems biology, or a model itself. Notably, not all ML predictions can be readily experimentally tested or validated. Thus, there will be a collaborative interplay between the need for testable predications and predictions that will in fact push cancer researchers to explore new experimental terrains.

Second, ML will provide the ability to analyze existing large, complex data sets to move from experimental hypotheses to guiding therapy, including identifying druggable targets, dosage strategies and quantifying uncertainty in response to therapy. Notably, due to the enormous size of data involved, the aforementioned cannot be done solely by experiment. One researcher recently estimated that for a single patient, this vast "therapy space" includes approximately 10^{40} possibilities[vi]. Applying ML on HPC platforms will enable the integration of large sets of molecular and visual data to strengthen research and diagnostic capabilities.

Innovation Potential for High-Performance Computing Capabilities

Meeting participants noted that applying ML to cancer challenges will push HPC and advanced computing in four key ways. First, it will drive advanced computing through the need for new ML models for synthetic, potentially multiscale, data generation (e.g., Generative Adversarial Networks [GAN]). Similarly, massive amounts of extremely heterogeneous data and varying metadata will require advances in modeling and data-integration capabilities. There will also be the need to develop new meta-rules for metadata to allow for integration of existing and new technologies into software/API/front-end to enable cancer biologists to easily add, edit and analyze their data.

Second, applying ML to cancer challenges will drive the development of new heterogeneous computational architectures, such as those needed for Deep Learning AI approaches. Thus, the application of ML to complex cancer data will provide a field for developing and testing advanced architectures on complex biological data.

Third, ML with cancer data will push the field of uncertainty quantification by supporting ML predictions with an associated measure of certainty and reducing input noise of large, heterogeneous real-world data. To be validated, ML outputs will need to be "explainable" to cancer clinicians and researchers. This will require a new level of algorithmic transparency and thus a deeper understanding of ML, especially in applications.

Fourth, this challenge will push the limits of reproducible and comparable data science. As new methods, architectures and datasets become available, the community will need to automatically generate new models and compare outputs to determine the best methods to use and combine. Part of generating new models will be automatic ML (AutoML), such as intelligent/optimized architecture search for both neural networks and ensemble stacking approaches.

Cancer Challenge Area 3

Creating Digital Twin Technology

"How can we simulate a cancer patient's care trajectory from pre-diagnosis to survivorship?"
-- Tina Hernandez-Boussard, Ph.D., Stanford University

Today, cancer care teams cannot offer patients a personalized view of their health trajectories, particularly when faced with various treatment options. Meeting participants envisioned a future in which a patient's digital twin (aka, avatar or virtual patient) could be used as a holistic *in-silico* model in cancer wet lab research, clinical trials and in clinical settings to guide more effective and personalized treatment choices. Digital twins would incorporate models of relevant biological processes, as well as disparate kinds of data unique to the patient. The technology would not just be used to stratify patients, but to predict the dynamics of their disease trajectories. This would expand precision medicine to *predictive* medicine.

Creating digital twin technology would be a grand challenge in HPC and oncology. It involves bridging spatiotemporal scales as never before—from the molecular, cellular, and tissue levels to the individual, population, and environmental levels. At each scale, agents interact with each other, and it will be necessary to identify the multitude of variables, many not currently captured systematically, that allow bridging and connecting of scales.

Many participants were enthusiastic about the potential of this systems-based approach and agreed that the digital twin is the ultimate multi-scale model. They also agreed that creating digital twin technology could only be accomplished through a dynamic, large-scale, multidisciplinary collaboration. As such, it is a major opportunity for HPC and oncology co-design efforts.

"That's a really big challenge: to connect many different models with different pieces and get something coherent out when you're done. So, it's a fascinating computer engineering, computer science and software mathematics problem that's going to take state-of-the-art computer facilities to actually build, train and explore these models."
-- Paul Macklin, Ph.D., Indiana University

Relevance

The creation of digital twins could completely alter basic, translational and clinical cancer research, treatment, and population health by providing an advanced, *in-silico* modeling environment across the oncology spectrum. Researchers and clinicians need to understand the inter-relationships of spatiotemporal scales, in both healthy and disease states, to predict the impact of molecularly targeted treatment for the individual, and how the individual's environment, behavior, etc. impacts molecular, cellular, and overall physical level response. The digital twin would provide researchers with a computational tool to formulate predictions based on hypotheses and approximations that would improve over time with recourse to finer-scale calculations and observations.

A digital twin would enable iterative and ensemble "what-if" evaluations of proposed interventions. This would enable physicians to not just better select the most effective treatment, but help patients weigh their treatment choices against their personal priorities and constraints. The digital twin population could identify high-risk populations and allow policymakers to evaluate different screening practices and guidelines. The digital twin capability has the potential to significantly impact policy and population health. In a clinical setting, a digital twin would also be a powerful tool for patient-physician communication to enable better informed patient choice and shared decision making.

Why Now?

This lean-in challenge builds on existing, but uncoordinated, efforts to create the first proto-digital twins. These pioneering models are far from being whole-patient representations. For example, German researchers are using a very rudimentary virtual model to select the best treatments for melanoma patients[vii]. There are also extensive examples of advanced computing models of individual cells and organs[viii].

Meeting participants noted that now is the time to harness new and emerging HPC resources to combine existing specific computational oncology models, such as those for tumor growth and vascularization, into a holistic, multi-scale model that can even produce population-level models. Notably, creating this integration with uncertainty-aware calculations requires massively parallel ensembles of simulations and analysis tasks that utilize emerging HPC systems. Collaborative efforts across disciplines are underway, and there is a need to coordinate efforts to deal with rapidly evolving data streams with various quality and time-scale issues[ix].

Innovation Potential for Cancer Research

"There's tremendous potential to drive innovation in cancer research at a variety of scales."
-- Paul Macklin, Ph.D. Indiana University

Digital twins promise to greatly increase resolution and decrease uncertainty in cancer research. A multi-scale framework will incorporate genomic, molecular, cellular, and population models that are consistent across space and time scales. These models can incorporate social, behavioral and environmental factors such as diet and pollution exposure.

One suggested biological framework for the digital twin-model presented by Meeting participants is the Hallmarks of Cancer[x]. These are defined as phenotypic changes at the cellular level that are shared by most, and possibly all, cancer types. However, these hallmarks of cancer are mostly studied in isolation and have proven to be of limited predictive utility at clinical scales. A digital twin program to computationally integrate these disparate hallmarks of cancer into one coherent model could be a major step toward understanding, predicting, and reducing cancer lethality.

At point of care, digital twins could provide personalized evidence to guide treatment decisions. Patients would be able to see their virtual twin across multiple treatment scenarios, providing personalized information of their cancer progression, treatment related side-effects, and quality of life.

The use of a population of digital twins, combined with leadership-class computing power, could augment the gold standard randomized clinical trial and enable rapid virtual clinical trials.

These might be able to quickly and efficiently identify potential treatment failures and opportunities. The time, resources and cost of conducting current clinical trials make this a compelling alternative, including potentially saving billions of dollars in the development of new drugs.

Innovation Potential for High-Performance Computing Capabilities

The step-wise development of digital twins would broadly drive innovation in both HPC architectures and advanced computing. The complexity of multi-scale, high-resolution, predictive models is anticipated to be more difficult than pure-physics models, thus pushing the state-of-the-art in HPC predictive science. For example, while most physics-based models involved proximate interactions over short time scales (femtoseconds to seconds), a digital twin would involve modeling across the time frames of molecular interaction to multiple years in a patient's life. Similarly, digital twin models must include cancer's ability to metastasize, thus, to act at a distance. Population models for virtual prevention trials would act over decades, over the entire space of the U.S.

The biological models will need to be validated and doing so will push the art of verifying models in complex systems. Similarly, a digital twin model would push the frontier of uncertainty quantification and error estimation that reflect both computational and oncological sources of error.

Cancer Challenge Area 4

Adaptive Treatments

Meeting participants noted that a key hurdle in the development and implementation of more effective cancer therapies is that cancer comprises many different diseases and is highly heterogeneous within tumors, between tumors and across patients. Tumors are a mix of heterogeneous cell types, can exist with dozens of slightly different genetic variants and can arise through clonal evolution. Cancers are mobile (metastasis), both infiltrating new tissue and triggering distal tissues to recruit cancer cells. Thus, cancer is enormously *adaptive*—hence the need for *adaptive* treatments.

The vision for adaptive treatments involves the development of biological and nano-device-based, personalized drug treatments that adapt to tumors over time. Creating these treatments would require the use of computational oncology. This approach builds on precision medicine, extending it to a new paradigm for cancer treatment. This challenge imagines direct-to-tumor interactive treatments that:

- Adapt to changing tumor characteristics during treatments;
- Target and attack metastasizing cancer cells, augmenting the immune system;
- Deliver novel therapeutics that fabricate molecules *in-vivo* at the tumor site.

"What we're talking about doing is programming a bacterium or programmable nano-device to change its behavior in a tumor micro-environment in response to the specifics of a patient. Think about a bacterium or nano-device as a mini, portable chemistry factory. Plus, the fact that it's going to synthesize the drug at the site, using materials that it has in the environment."

-- Rick Stevens, Argonne National Laboratory

Relevance

The systemic adaptation and multi-variant behavior of cancer must be addressed in future-generation therapies. At present, this is sometimes successful—but often only temporarily—through a sequential or concurrent combination of radiation, chemotherapy, surgery, immunotherapy, proton therapy and other treatments. However, for many cancer sub-types and late-stage cancers, these therapies are largely ineffective. Thus, there is the need to imagine, develop, test and implement a concurrently robust response: adaptive therapies.

Why Now?

This lean-in challenge area would not have been addressable ten years ago. At present, however, there is a confluence of theoretical, modelling and applied sciences, infrastructure and tools that offer the possibility to imagine and create adaptive treatments. These tools include synthetic biology, systems biology, genetic engineering tools (ex. CRISPR), ML, nanofabrication, nanorobotics, simulation, and clinical applications in bacterial and viral oncology therapies.

Synthetic biology centers on the design, construction, and characterization of improved or novel biological systems using engineering design principles. At present, the United States is the world-leader in synthetic biology and there have been several national level roadmap exercises[xi].

Innovation Potential for Cancer Research

This lean-in challenge of Adaptive Treatments leverages and extends, a resurgence of interest in natural bacterial approaches to cancer therapy[xii]. Spontaneous tumor regression has been associated with microbial infection for hundreds of years and, a century ago, inspired American physician, William Coley, M.D. (1862–1936), to pioneer the use of live bacteria as a deliberate cancer treatment. In the past ten years, progress has been made with a variety of bacterial organisms, treating a variety of cancers in cell lines, model systems and a few clinical trials. Similarly, viruses (phage) have been used as experimental oncological treatments. The limited success of these approaches demonstrates potential utility[xiii].

This lean-in challenge extends these pioneering bacteria-based approaches by applying computationally driven synthetic biology to the engineering of adaptive biologics and nano-devices. The therapies would monitor the tumor as it develops and respond accordingly to destroy cancer cells. This synthetic bacterial treatment would be a combination sensor (diagnosis) and adaptive treatment factory, able to synthesize dozens of cancer-fighting molecules in response to sensor data. Bacteria and envisioned nano-devices are very small compared to tumor cells and could be used to augment the immune system to destroy metastasizing cells.

Innovation Potential for High-Performance Computing Capabilities

> *"Engineering this (creation of adaptive treatments) requires the ability to first imagine how it might work, get it working in simulation while we try to get it working in an actual system and do essentially exhaustive virtual clinical trials."*
>
> -- Rick Stevens, Argonne National Laboratory

Developing adaptive therapies provides a next-generation challenge for HPC and advanced computing. The first key challenge will be to integrate and scale existing approaches in synthetic biology. This would push both HPC architectures and advanced computing. On a socio-cultural and political level, infecting patients with a synthetic pathogen as a means of cancer treatment might well be a controversial issue both at the patient and physician level. Thus, advanced modeling, simulation and ML will play a key role prior to testing treatments *in vivo* and for guiding *in vitro* and *in vivo* research.

Modeling and ML will drive synthetic biology discovery by using Deep Learning to define the relationship between the tumor properties and the pathogen properties to infect the tumor and the variation within it.

---

[i] Hinkson, I. V., Davidsen, T. M., Klemm, J. D., Kerlavage, A. R., & Kibbe, W. A. (2017). A Comprehensive Infrastructure for Big Data in Cancer Research: Accelerating Cancer Research and Precision Medicine. *Frontiers in cell and developmental biology*, 5, 83. doi:10.3389/fcell.2017.00083

[ii] Johnson AEW, Pollard TJ, Shen L, Lehman L, Feng M, Ghassemi M, Moody B, Szolovits P, Celi LA, Mark RG. (2016). MIMIC-III, a freely accessible critical care database. *Scientific Data*. *3*, 160035. doi: 10.1038/sdata.2016.35

[iii] Robert L. Grossman, Allison P. Heath, Vincent Ferretti, Harold E. Varmus, Douglas R. Lowy, Warren A. Kibbe, Louis M. Staudt. 2016. Toward a Shared Vision for Cancer Genomic Data. *N Engl J Med*, *375*,1109-1112. doi: 10.1056/NEJMp1607591

[iv] Gerard J. Kleywegt, Sameer Velankar, Ardan Patwardhan. (2018). Structural biology data archiving–where we are and what lies ahead. *FEBS Letters*, *592*,12, 2153-2167. doi.org/10.1002/1873-3468.13086

[v] Derek Wong & Stephen Yip. (2018). Machine Learning Classifies Cancer. *Nature* 555, 446-447. doi: 10.1038/d41586-018-02881-7

[vi] Berry, Donald A., (2015). The Brave New World of clinical cancer research: Adaptive biomarker-driven trials integrating clinical practice with clinical research. *Molecular Oncology*, *9*, 951-959. doi: 10.1016/j.molonc.2015.02.011.

[vii] "Need a Doctor? Send in your digital twin," *Cancerworld*, Sept. 24, 2018. (Retrieved: May 20, 2019 from https://cancerworld.net/spotlight-on/need-a-doctor-send-in-your-digital-twin/)

[viii] Cranford, J.P., O'Hara, T.J., Villongco, C.T. et al. (2018). "Efficient Computational Modeling of Human Ventricular Activation and Its Electrocardiographic Representation: A Sensitivity Study. Cardiovasc Eng. Tech. *9*: 447–467. doi.org/10.1007/s13239-018-0347-0

[ix] J. Ozik, N. Collier, R. Heiland, G. An, & P. Macklin. 2019. Learning-accelerated Discovery of Immune-Tumor Interactions. *Molec. Sys. Design Eng.* (*in review*). Preprint: https://dx.doi.org/10.1101/573972

[x] Hanahan, Douglas et al. (2000). The Hallmarks of Cancer. *Cell*, *100*,1:57-70.

[xi] Si, T., & Zhao, H. (2016). A brief overview of synthetic biology research programs and roadmap studies in the United States. *Synthetic and systems biotechnology*, *1*(4), 258–264. doi:10.1016/j.synbio.2016.08.003

[xii] Kramer MG, Masner M, Ferreira FA and Hoffman RM (2018) Bacterial Therapy of Cancer: Promises, Limitations, and Insights for Future Directions. *Front. Microbiol*. *9*:16. doi: 10.3389/fmicb.2018.00016

[xiii] Ibid.