NCI-DOE Digital Twin Micro Lab, April 23, 1:00 p.m.-2:30 p.m. ET

120 people attended the Micro Lab. Emily Greenspan, Center for Biomedical Informatics and Information Technology, National Cancer Institute; Sean Hanlon, Center for Strategic Scientific Initiatives, National Cancer Institute; and Eric Stahlberg (Frederick National Laboratory for Cancer Research) provided an overview of the NCI-DOE Collaboration and previous activities that led to the Micro Lab and introduced the upcoming Ideas Lab, scheduled for July 6-10, 2020.

Breakout Session Notes

After the opening presentations, 48 people participated in 9 breakout discussion groups. Below are the questions they were tasked with and a comprehensive list of responses.

1. What might a successful digital twin look like?

- **How data might be handled in a digital twin**
  - Connecting data across scales.
  - Construct from a molecular database of profiles, look at therapies that were attempted and any tumor response/resistance.
  - Twins are generated from many samples, can aggregate samples and look at trends.
  - Studying people longitudinally is hard because getting samples over time is hard. Liquid biopsies are promising to monitor over time and have empirical evidence of early resistance/response through blood.
- **Predictive tool**
  - Using germline mutations to predict cancer occurrence, e.g., when BC vs. CRC? Goal: combine multi-omics data from electronic health records (EHR) and data from other patients.
  - A tool that can give insight into how a treatment will be useful with rationale of why. To get a response in a reasonable amount of time.
  - Digital data is epigenetics, molecular in digital archive (digital representation along with responses to different therapies).
  - A reasonably predictive model that can intake personal history and general omics data, clinical measurements, imaging data, and information from personal wearables and/or environmental information to in a reasonably quick time provide predictions that can lead to a) testing hypotheses and design of new experiments and b) ultimately inform physicians of outcome of different treatments. We have to a user (healthcare provider and patient) centered design.
  - Multi-scale (organ and tissue level, molecular, EHR scale, environmental). Dynamical models that can make predictions for the patient in the future. Make a personalized digital human (we want to not just focus on cancer) so that we

are taking the whole person into account (including comorbidities and environment). And the model will be of the whole patient and will likewise not just predict survival or response to a treatment. It will be on a much larger scale.
    - A model that predicts application of treatments, is adaptable to the variance of global population, can be validated and verified, shareable, improvable, standardized, and maintains appropriate level of patient privacy.
    - Be able to predict some finite time into the future; the further the better; state of individual and response to intervention.
- **Simulations that model cancer**
    - We can build on top of existing models (PDX, etc.), how therapies work and don't work from a molecular model system.
    - Patients with cancer have different phenotypes of their disease. Can radiology reports help inform clinical trials stratification? Currently, genomics may also be included.
    - Comes from particle physics. In cancer, don't have the model but have an incredible amount of data. How do we model parts and assemble this into a larger model? Very exciting and lots of interesting computer problems.
    - Over time you could look for patients where histopathology are similar and then aggregate, run simulations, look at tumor more robustly.
    - Can the mouse model match a digital twin generated from pure digital data? Use gene editing to make matching mice.

2. What could be some first steps toward creating a digital twin?

- **Have a specific use case/research question**
- **Compile data**
    - Identify the relevant tests to perform to compile the best data, including longitudinal, for analysis.
    - Identify the issues with data acquisition.
    - Make sure that new data from new technology is accessible. EHR suffers from missing data since patients see different doctors at different locations. Which data is most "trustworthy" for interpretation?
    - How do we collect the breadth of responses for a given patient? People respond in so many different aspects - how do we capture this?
    - Gather the data needed to model from the widest range of population to assure it is not class limited.
    - Utilize wearable device data and make them more accessible.
    - Education of patients on importance of data gathering and potential to be gained.
    - Data description as a precursor to a common data model.
    - Choose one potential user group (physicians) and a type of cancer that has a multi-modal and longitudinal with different time-frames data set. Choosing level

of models can be guided by physicians (cellular/biochemistry, pathway, organ/body). What are needs for HPC? Defining experiments. Improving data collection methodologies. What are the significant variables to include in a predictive model? How do we tune these to an individual?

- **Organize data**
  - Ontology
  - This is a big gap in the radiology space. As a clinician, treatment decision is key. RESIST is incompatible with daily clinical practice. Need to define response using Big Data.
  - How do you bring all the data together to do the model development?
  - What's a good dataset that's annotated for AI? This may be a good first step.
  - Harkens to uncertainty quantification. What can we learn about the uncertainty of the prediction? But also, can consider the quality of the incoming data. How about comparing data coming from different groups?
  - Cataloguing/evaluating available data and data sources (claims data, VA data, SEER, TCGA, HTAN, ASCO CancerLinQ®, etc.)
- **Establish a common vocabulary**
  - How do we agree on the variability in response in clinical practice? If we can agree on what progress means, this will be important progress. Question: how do you validate predictions of response? Can't do this patient by patient but can do it with Big Data. If there were a centralized database to compare to, a given radiologist could see how their reads compare to the rest of the Nation.
- **Use prior models as starting point**
  - Figure out what are the physical processes/biological systems that have already been modeled (and how well they work), identify the gaps and then identify what processes/systems still need to be modeled and how should the systems interconnect. Data collection of various sorts must involve models, and the parameters of these models could be determined with the help of machine learning.
  - Cataloguing/evaluating/benchmarking the existing models (both supervised and unsupervised) across scales to think about how to integrate.
  - Create robust mouse model with extensive molecular profiles.
  - Map genetic parameters to kinetic parameters (pre/post bloods for liquid biopsies).
  - Lung cancer and melanoma have relevant data, enough events in those cancers and relevant mouse models.
  - Could pick pancreas cancer because the genomic story is different and heterogeneous, stromal involvement.

    - Develop a framework for formulating mathematical models that will serve as the initial input into HPC servers. The models need to be modular and be able to import available models.
- **Start with organoids**
- **Landscape assessment/horizon scanning**

3. What might be some of the intermediary goals toward achieving a digital twin? What's the Moonshot to the digital twin Mars-shot?

- **Incremental moonshot**
    - Digital twin of different parts of the body
    - Digital twin of select groups of the population
    - Build digital twin of only newborns and begin to build on this.
    - Combining small numbers of models (across scales) that were validated in first step
    - Shorter term predictions
    - Focus on cancers where we know more or are less complicated.
    - Smaller piece within a diagnosis > digital tumor
    - Starting with a process that is similar across all/many tumor types
    - Model normal blood development process > then add in things that go wrong in various blood cancers.
- **Define goals and assumptions**
    - Need to define the first step.
    - Define digital twin—what should be useful?
    - How should we answer the various questions; data/tools?
- **Link/share data**
    - Get institutions to share data in a way that is safe and beneficial for researchers and patients.
    - Variety of data being brought together for analysis (3 types of data)
    - Having all data linked as in France. Any doctor can access all the data on a given patient. Need good protection of the data and appropriate policies in place.
    - Could there be a nationalized database that patients can consent to? Patients say, "you can share my data with this nationalized body." No one researcher has enough data.
    - Aggregating consistent, well annotated data.
    - Identify the data sources and meta-data (context, reproducible research, public/private, additional co-morbidity data sets), establish collaboration with physician to develop access to use-case data and more rapid (agile) feedback for development, develop informed consent process. Develop more efficient data collection.

- **Craft/share models**
    - Another element of the knowledge base is hosting and sharing the models, that is continuously updated. Goes hand in hand with collecting the data. Also encourages different parts of the model coming together toward answering the larger question (like in particle physics).
    - Build models from fine detail to high-level processes so the models can be integrated.
    - Combine coarse and fine-grained models (across many orders of magnitude across time and space).
    - In atmospheric modeling this was done as a community effort, needs to be mission driven by a sponsor (NCI, DOD), sponsor can set standards, funding and road map/clearinghouse to organize questions.
    - Develop a PDX and treat with a certain therapeutic regimen and see if you simulate getting from point A to point B (treated to untreated), a well-characterized model system to drive prototyping (use lung cancer - EGFR with a point mutation, let the mouse grow a little and then molecular profile [point A], let the rest of the vehicles grow [point B], and then on to point C.
    - 1. Validate models from the use case. 2. Broaden the model to account for more use cases. The moon shot would be a pan-cancer model that would be able to identify disease probabilities, diagnose and predict the efficacy and safety of a treatment.
    - Figure out what assumptions do we make so that we don't have to measure too much so that the model is still predictive. Collect models of the systems first. Models could be stochastic; exploration of the parameter space should also be done. Model the cells, the tissues, the organs and then figure out how to interconnect these different systems. Determine which features are most important to determine the parameters for the model. What is the fidelity of the models relative to the biological processes that they are modeling (i.e., how predictive are these models?). Must be determined by the desired outcomes from these models.

4. What are the barriers to creating a digital twin?

- **Access to data/lack of data**
    - The spread of data across multiple platforms and access to the data
    - Lack of number of samples (not enough to make reasonable predictions)
    - Data generated not for data science but tracking of Medicolegal and clinical documentation
    - Data access (the "right data" for a model - what can we measure to inform the model? Omics may not be the most useful source that addresses all aspects of a dynamical model—hard to get longitudinal data)
    - Data management/organization

  - Integration of private and public data
  - Accurate, consistent data
  - Availability of data
  - Time-dependent data
  - May be impossible to start with old data
  - Collection of contextual data
  - How much data do you need? What kinds of data do you need?
  - How will missing data impact output?
  - Lack of data from some populations
- **Knowledge gaps**
- **Data privacy**
  - Patient privacy concerns
  - PII/HIPPA protection, patient consent. How do we test a digital twin model, quantify uncertainty (FDA conversation)?
- **Trust**
  - In France, only doctors and pharma can access the data. Can use this information against the patient. This makes a big difference. Important to trust who manages the data. Need both trust and technology.
  - Intellectual property is a big one. Who as disclosures they are not disclosing? Who is making money? Need a national, trusted authority. So difficult in the US.
    - Example: 23 and Me, people are paying to give away their data if they are motivated. Bypass lawyers and let patients collaborate with each other.
- **Lack of models/model accuracy/integration of models**
  - Predictable model accuracy
  - Connecting the heterogeneous (multi-scale models) together will be a significant challenge, also taking into consideration that many of the models are at different time scales. Initially getting participation of patients may be hard but will eventually get easier as the project(s) progress and patients see benefits.
  - Ability to validate and verify the results of the model
  - Interaction between model and stakeholders
  - Ensemble of many models may compound the challenges
  - Lack of integration of different types of mathematical models that account for different biological processes
- **Creation of class disparity with only select individuals having access to this information**
- **Uncertainty/accuracy in prediction**
  - How far into the future is prediction feasible?
  - How accurate is prediction?
- **Knowing when models are good enough and can be trusted**
  - Are the models and components good enough?

- Sensitivity? What can you believe and what needs to be questioned?
- Trust of model? Do they capture biases?

Micro Lab Attendees by Affiliation

Ansys	1
Booz Allen Hamilton	1
Brigham and Women's Hospital, Harvard University	1
Brookhaven National Laboratory	1
DOE Office of Science	2
Duke University	1
Food and Drug Administration	1
Frederick National Laboratory for Cancer Research	24
GE	3
Health Partners	4
IBM	1
Icahn School of Medicine Mount Sinai	1
Indiana University	2
Instituto Nacional de Astrofisica Optica Y Electronica, Mexico	1
Intel	1
IPQ Analytics	1
Johns Hopkins University	3
Lawrence Livermore National Laboratory	6
Los Alamos National Laboratory	4
Mayo Clinic	2

Memorial Sloan Kettering Cancer Center	1
NCI/NIH	26
Oakridge National Laboratory	1
Pacific Northwest National Laboratory	2
Partners Healthcare	1
Purdue University	2
Queens University	3
SpIntellx	1
Institute for Systems Biology	2
TGen	1
University of California, Santa Barbara	1
Unity Health Group	1
University of Arkansas for Medical Science	1
University of California, Davis	1
University of Chicago	1
University of Maryland, Baltimore City	1
University of Michigan	1
University of Minnesota	1
University of Texas at El Paso	1
Washington University in St. Louis	2
Unknown	3

Micro Lab Breakout Groups

Group 1
Jon Mohl, University of Texas at El Paso
George Zaki, Frederick National Laboratory for Cancer Research
Hannah Lonsdale, Johns Hopkins University
Prabhu Shankar, University of California, Davis
Tanveer Syeda-Mahmood, IBM Research
Amit Mistry, National Institutes of Health

Group 2
Juli Klemm, National Cancer Institute
Tiphaine Martin, Icahn School of Medicine at Mount Sinai
Richard Do, Memorial Sloan Kettering Cancer Center
Eric Stahlberg, Frederick National Laboratory for Cancer Research
Michael Cooke, Department of Energy

Group 3
Amber Simpson, Queen's University
Paul Macklin, Indiana University
Sarangan Ravichandran, Frederick National Laboratory for Cancer Research
Don Johann, University of Arkansas for Medical Science
Bill Cannon, Pacific Northwest National Laboratory

Group 4
Emily Greenspan, National Cancer Institute
Brion Sarachan, General Electric
Ella Saccon, National Cancer Institute
Judith Cohn, Los Alamos National Laboratory
Ali Navid, Lawrence Livermore National Laboratory
Patricia Mabry, HealthPartners Institute

Group 5
Naomi Ohashi, Frederick National Laboratory for Cancer Research
Marian Anghel, Los Alamos National Laboratory
Brian Luke, Frederick National Laboratory for Cancer Research
Taina Immonen, Frederick National Laboratory for Cancer Research
Jack Collins, Frederick National Laboratory for Cancer Research

Group 6
Ilya Shmulevich, Institute for Systems Biology
Samuel O'Blenes, HealthPartners Institute
Zorina Galis, National Institutes of Health - National Heart, Lung, and Blood Institute
Anupama Govindarajan, Food and Drug Administration
Nadia Lanman, Purdue University

Group 7
Amar Khalsa, Frederick National Laboratory for Cancer Research
Michael Tschanz
Jaume Reventos
Karen Batch, Queen's University
Yue Dong, Mayo Clinic

Group 8
Sean Hanlon, National Cancer Institute
Jennifer Couch, National Cancer Institute
Frank Alexander, Brookhaven National Laboratory
Sanjay Purushotham, University of Maryland Baltimore County
Pamala Pawloski, HealthPartners Institute
Ganesh Chand, Washington University in St. Louis

Group 9
Kirsten Meeker, University of California, Santa Barbara
Timothy Fries, Booz Allen Hamilton
Lynn Borkon, Frederick National Laboratory for Cancer Research
Amy Gryshuk, Lawrence Livermore National Laboratory
Boris Aguilar, Institute for Systems Biology