U.S. Department of Health and Human Services
Public Health Service
National Institutes of Health
National Cancer Institute

15th Virtual Meeting
Frederick National Laboratory Advisory Committee

Summary of Meeting
11 March 2024

National Cancer Institute
National Institutes of Health
Bethesda, Maryland

National Cancer Institute
15th Virtual Meeting of the Frederick National Laboratory Advisory Committee

11 March 2024

Summary of Meeting

The Frederick National Laboratory Advisory Committee (FNLAC) convened for its 15th Virtual Meeting on 11 March 2024. The meeting was open to the public from 1:00 to 4:11 p.m. EDT. The FNLAC Chairperson, Dr. Candace S. Johnson, President and CEO, M&T Bank Presidential Chair in Leadership, Roswell Park Comprehensive Cancer Center, presided.

FNLAC Members
Dr. Candace S. Johnson (Chair)
Dr. Carol J. Bult
Dr. John H. Bushweller
Dr. Timothy A. Chan
Dr. Lisa M. Coussens (absent)
Dr. Blossom A. Damania
Ms. Julie Papanek Grant (absent)
Dr. Angela M. Gronenborn
Dr. Mary J.C. Hendrix
Dr. Rodney J.Y. Ho
Dr. Allison Hubel
Dr. Dineo Khabele
Dr. Anant Madabhushi
Dr. Patrick Nana-Sinkam
Dr. Nilsa C. Ramirez Milan
Dr. Erle S. Robertson
Dr. Linda F. van Dyk

NCI Senior Leadership
Dr. James H. Doroshow
Dr. Anthony Kerlavage
Dr. Kristin L. Komschlies
Ms. Anne Lubenow
Dr. W. Kimryn Rathmell
Dr. Dinah S. Singer

Executive Secretary
Dr. Christopher D. Kane

TABLE OF CONTENTS

I. OPENING REMARKS—DR. CANDACE S. JOHNSON

Dr. Candace S. Johnson, Chair, called to order the 15th Virtual Meeting of the Frederick National Laboratory Advisory Committee (FNLAC) and welcomed the Committee members, National Cancer Institute (NCI) staff, and guests. Dr. Johnson reminded members of the conflict-of-interest guidelines and confidentiality requirements. Members of the public were welcomed and invited to submit to Dr. Christopher D. Kane, Executive Secretary, in writing and within 10 days, any comments regarding items discussed during the meeting.

Motion. A motion to approve the minutes of the 19 October 2023 FNLAC meeting was approved unanimously.

Dr. Johnson called the Committee members' attention to the confirmed future meeting dates listed on the agenda, noting that the next FNLAC meeting will be held on 9–10 July 2024 and is planned to be held as an in-person meeting.

II. NCI DIRECTOR'S REPORT—DR. W. KIMRYN RATHMELL

Dr. W. Kimryn Rathmell, Director, NCI, also welcomed the FNLAC members and attendees to the meeting. She provided a budget update and reported on NCI news, as well as research highlights from the Frederick National Laboratory for Cancer Research (FNLCR). Dr. Rathmell began by describing her clinical and research background and emphasized that she values listening and openness, teamwork and collaboration, and solving complex problems. She noted that her laboratory was involved in The Cancer Genome Atlas Program, which relies on resources from the FNLCR. She emphasized that this program underscores the value of federally funded research and development centers, such as the FNLCR.

NCI Fiscal Year (FY) 2024 Budget. Dr. Rathmell noted that a government shutdown due to lapses in appropriations has been averted multiple times in FY 2024 with three continuing resolutions (CRs): 30 September 2023, 17 November 2023, and 2 February 2024. The current CR expires 22 March 2024, which is halfway through the fiscal year. NCI has been developing a set of interim grant policies, and final adjustments will be made when the appropriations are finalized. Dr. Rathmell explained that the appropriations bills for NCI's budget are grouped with those of other federal agencies, including the U.S. Department of Homeland Security and U.S. Department of Defense. Dr. Rathmell emphasized that she has engaged with members of Congress to convey the importance of NCI's work.

Dr. Rathmell reviewed how NCI spends its appropriations, noting that 74 percent of the NCI budget supports extramural grants; the majority of these are research project grants (RPGs), which are traditionally R01s. Of the NCI budget, 8 percent supports NCI-Designated Cancer Centers (Cancer Centers) and Specialized Programs of Research Excellence (or SPOREs). The remainder supports intramural research, research management and support, and buildings and facilities. Further details can be found on the NCI Budget and Appropriations website. Dr. Rathmell emphasized that the FNLCR is a critical component within NCI's budget process.

NCI's FY 2023 budget was $7.3 billion (B). Per the National Cancer Act of 1971, NCI submits a Professional Judgment Budget Proposal (also called the Bypass Budget) directly to Congress. This Bypass Budget estimates the cost of the work that NCI is expected to perform. The overall cost of inflation has been increasing more than the budget estimates for several years. Already behind, NCI has been asked to implement new initiatives and aims to improve cancer health care. The *Annual Plan and Budget Proposal for Fiscal Year 2024* was increased to $9.9 B, which the President considered before recommending a budget of $7.8 B. NCI remains optimistic about the FY 2024 budget and its approval. FY 2025 begins 1 October 2024, and for the 2025 Professional Judgment Budget, NCI is proposing a

budget increase to $11.5 B, which reflects inflation and projected costs for such efforts as conducting clinical trials; gathering and analyzing cancer data; and using the data to better serve communities. The President's budget proposal will be released in spring 2024. Dr. Rathmell noted that the Bypass Budget highlighted several scientific opportunities, including improving patients' lives through symptom science research, revolutionizing cancer clinical trials, clarifying the impact of the environment on cancer risk, harnessing the power of cancer data, and unraveling the complexity of cancer metastasis.

Dr. Rathmell explained that NCI has been projecting the implication of a "flat" FY 2024 budget, noting that it will be challenging to maintain current efforts. Expenses will increase because of inflation and staff salary increases. NCI incurs between $75 million (M) and $100 M annually in increased mandatory expenses (e.g., program evaluations, cybersecurity, Center for Scientific Review expenses). To maintain the 12th percentile payline, NCI will need to add $250 M to the RPG pool to fund both new and noncompeting awards at 100 percent. Without the additional $250 M, NCI would be confronted with decreasing the payline for new RPG awards and funding noncompeting awards at less than 100 percent. Decreases that were not previously considered will need to be contemplated for the competing renewals of Cancer Center Support Grants (CCSGs) and cancer training awards, as well as for the noncompeting CCSGs. Cuts to the intramural research program are anticipated to be along the same levels as to the extramural awards. Additional information can be found on the NCI website, including the *NCI Bottom Line: A Blog About Grants and More*.

Recent News and Events. Dr. Rathmell highlighted several NCI updates. The National Cancer Plan (NCP) was implemented in 2023. The eight goals of this roadmap encompass the need to prevent cancer, detect cancers early, develop effective treatments, deliver optimal care, maximize data utility, eliminate inequalities, optimize the workforce, and engage every person. The NCP is focused on changing how we think about cancer, reducing cancer mortality, and improving the lives of people with cancer. The NCP aligns closely with the Cancer Moonshot℠ and recently was reviewed by the President's Cancer Panel, which assessed each goal and presented a series of recommendations. President Joseph R. Biden announced the reignited Cancer Moonshot 2 years ago, setting bold but achievable goals to reduce the U.S. cancer death rate by 50 percent by 2047 and to improve the experience of patients with cancer and their families.

Research Highlights. Dr. Rathmell presented recent successes in cancer research, with a focus on projects that are relevant to the FNLCR. She noted that the RAS Initiative recently developed a first-in-class direct inhibitor of KRAS G12C. RAS genes are involved in cell growth, cell maturation, and cell death and are mutated in more than 30 percent of cancers. KRAS is one of three human RAS genes. She also noted that results of the first randomized controlled clinical trials on single-dose human papillomavirus vaccine efficacy have been performed. NCI also is pursuing big data initiatives (e.g., use of artificial intelligence [AI] for rapid diagnosis of endometrial cancer); these efforts are aligned with the NCP's goal to maximize data utility. Dr. Rathmell remarked that these efforts provide researchers the opportunity to examine an underserved cancer and address gaps that relate to these disparities.

NCI recently formed new collaborations on antibody and drug conjugates, and the FNLCR can play a role in developing new treatments with reduced side effects. Dr. Rathmell also remarked that the FNLCR is partnering with the American Cancer Society and Cancer Research UK to organize the first U.S.-based Black in Cancer meeting, which will be held 20–21 June 2024. She underscored the importance of a diverse talent pool and an ongoing developed pipeline among cancer researchers. She reminded the FNLAC members that NCI is a part of the Cancer Grand Challenges, which demonstrate the power of team science. The four recently funded challenges are reducing cancer inequities, understanding mechanisms of early onset cancers, developing drugs for solid tumors in children, and broadening knowledge of how T cells recognize cancer cells.

In closing, Dr. Rathmell emphasized the importance of good science, rigor, honesty, and thoughtful prioritization during challenging times. NCI will use the NCP as the roadmap to achieve the Cancer Moonshot goals of ending cancer as we know it. She expressed appreciation to the FNLAC members for their insight and forward thinking.

In the discussion, the following points were made:

- Information on NCI's recent activities (e.g., endometrial cancer diagnosis, Black in Cancer meetings) can be found on NCI's website, but further efforts toward dissemination (e.g., social media) would help enhance their impact. Additionally, NCI could consider ways to promote training and interdisciplinary work within the Black in Cancer program.

- The R01 grants are increasing at a similar rate to the grants overall. The Advanced Research Projects Agency for Health (ARPA-H) grants have not yet affected the NCI portfolio, but these activities have encouraged the research community to consider what types of projects are ready for this type of mechanism.

- Projects associated with the Cancer Grand Challenges potentially could be applied to activities at the FNLCR.

- Multiple opportunities for public–private partnerships are available in the space of AI research; substantial progress has been made in this field.

III. DCTD'S BIOPHARMACEUTICAL DEVELOPMENT AND PRODUCTION AT THE FREDERICK NATIONAL LABORATORY FOR CANCER RESEARCH (FNLCR)—DR. JASON YOVANDICH

Dr. Jason Yovandich, Chief, Biological Resources Branch, Biopharmaceutical Development Program (BDP), Division of Cancer Treatment and Diagnosis (DCTD), NCI, presented updates on FNLCR's biopharmaceutical development and production. He explained that the division aims to provide resources to the community, fill gaps, foster development of new biotechnologies, disseminate the knowledge gained, and share unique resources with NIH partners. The division is focused on developing biological isolates and extracts, recombinant proteins, antibodies, antibody–drug conjugates and associated imaging agents, virus products and virus-like particles, and bacterial-based therapies. Newer areas of focus include engineered cells and synthetic biology.

The DCTD applies due diligence and preclinical testing of the material to analyze the starting materials for Good Manufacturing Practices (GMP) suitability. This involves measurements of safety, identity, strength, quality, and purity (SISQP). The process development stage determines scale-up feasibility, purification development, assay development for standard operating procedures (SOPs), and initial formulation stability. The next step is to generate sufficient reference material for the biologic candidate, source starting and production materials, and generate cell banks and/or virus banks. Next, the quality control group establishes analytical qualifications for product release. Pilot manufacturing is performed to produce material that can be used for reference standards, as well as for toxicology and infusion apparatus stability studies. The final steps include manufacturing the clinical lot and applying real-time stability. At some point, preferably prior to the tox studies, the team engages with the U.S. Food and Drug Administration (FDA) to discuss plans for Investigational New Drug (IND) filing. New projects are evaluated via standard questionnaires to gauge the developmental readiness of the agent. The questionnaire addresses SISQP and contains specific questions for biologics that pertain to recombinant proteins, viruses, peptides, and cell products.

Dr. Yovandich explained that the BDP was established by DCTD in the 1990s at the FNLCR to provide specialized and unique technical expertise that is not readily available in the commercial market. These activities are aligned with the laboratory's goals and mission as a federally funded research and development center. The program conducts feasibility studies on project candidates, develops manufacturing processes and assays, performs the GMP manufacturing, and produces and submits documentation required for FDA and international regulatory filings. The program has a streamlined process for technology transfer as commercial entities take over the product for further development, and a GMP training program is in place for staff and external partners. Overall, the program has produced more than 260 investigational agents, including recombinant proteins or peptides, cellular products, monoclonal antibodies, viral vectors, and virus-like particles. Dr. Yovandich briefly shared an overview of the facility layout and features, which include a GMP manufacturing area, self-contained virus manufacturing suite, cell therapy suites, and spaces for process development, analytical testing, and for warehousing products and raw materials. Equipment in the space includes bioreactors, fermenters, ÄKTA chromatography skids, multichannel cell sorters, high-performance liquid chromatography analyzers, and a semiautomatic filling machine.

The BDP follows the U.S. Code of Federal Regulations that pertains to current GMP practices for finished pharmaceuticals. Because the program focuses on early-stage development, it has established a quality manual that describes how the program will adhere to regulations at the level that is phase appropriate. Personnel undergo a specialized training curriculum based on their department, job function, and responsibilities. SOPs for documentation are in place across the program, including manufacturing, quality control testing, facility operations, equipment operations, and environmental monitoring. A validated electronic document management system holds all SOPs, training records, and master production records. All utilities and sterilization equipment are validated and undergo recertification annually. Quantitative instrumentation is calibrated, and product contact equipment is validated. A risk-based approach is followed for product safety and quality impact assessment.

Recently, the BDP set up capabilities to develop and manufacture autologous chimeric antigen receptor (CAR) T-cell products. NCI is supporting two multisite clinical trials. Process validations and qualifications have been established. Through this effort, the program can turn around a cell therapy product within about 3 weeks from receipt of the starting apheresis material. The program also established procedures for shipping and product chain logistics that are capable of supporting multicenter trials, with the BDP being the central manufacturing site. The program has established platforms for producing GMP-grade lentivirus and gamma retrovirus that are used for engineering the cell products. Additionally, the program has established capacities for CRISPR-based cell engineering, with up to 40 percent gene knock-in efficiency.

Dr. Yovandich highlighted successful commercialization and product licensures resulting from the BDP. Lumoxiti, an anti-CD22 immunotoxin, was BDP's first commercialized product but has been pulled from the market. Unituxin, an anti-GD2 antibody for pediatric neuroblastoma, is currently an active commercial product. Lerapolturev, or PVSRIPO, is in the late stages of commercialization for treatment of recurrent glioblastoma, melanoma, and non–muscle invasive bladder cancer. Chimeric 11-1F4, also known as anselamimab, also is in the late stages of development. The program is supporting the cell therapy community and the Cancer Adoptive Cell Therapy Network (Can-ACT). When the program established its cell therapy manufacturing platform and capability, one of the first tasks was to support a multisite clinical trial for relapsed or refractory acute myeloid lymphoma (AML) in young adults and children. To date, the program has produced 26 products successfully and has supported the clinical trial up to the maximum tolerated dose of the products. The DCTD has also sponsored a multisite trial through the Pediatric Early Phase Clinical Trials Network (PEP-CTN), for which BDP produces autologous CART products.

The BDP supports DCTD's Can-ACT network, which is fostering innovation and early stage clinical testing of novel, state-of-the-art, cell-based immunotherapies for solid tumors in adults and pediatric patients. Can-ACT's goals are to develop and enhance cellular products, through either genetic modification or other manipulations; support early phase clinical trials; explore imaging and biomarker development; expand the understanding of the mechanisms of action, as well as resistance mechanisms that are encountered; and evaluate strategies to modulate the immunosuppressive tumor microenvironment. The network is still being established and includes four adult UG3 grants and a U24 coordinating center. It is anticipated that one to two pediatric UG3 grants will come onboard soon. A series of supplements is anticipated to bring P30/P50 cancer center affiliates into the network.

Of the four funded UG3 grants, three receive support from the FNLCR. Dr. Yovandich highlighted research projects supported by FNLCR's Immune Cell Network (ICN) Core: (1) A CAR T-cell strategy that targets mesothelin and expresses a dual agent that targets a fibroblast-activating protein and CD3 to recruit bystander T cells; (2) a T-cell receptor strategy targeting the G12V mutant KRAS using CRISPR-based editing; and (3) a CAR-targeting mesothelin for mesothelioma. The BDP task areas that are part of this ICN Core include guidance on quality systems and regulatory affairs—including the development and standardization of assays for determining product-critical quality attributes, providing reagents and SOPs to the network, and providing regulatory guidance—as well as GMP production for multisite trials including viral vectors, cell products, and novel production technologies.

Access to the BDP by the extramural community is provided primarily through the NCI Experimental Therapeutics (NExT) Program. NExT offers a free consultation service to anybody who is in the early stages of IND planning. The program has collaborated with various entities, including other NCI Divisions and Centers, the National Center for Advancing Translational Sciences, and the National Institute of Allergy and Infectious Diseases. Other partners include the National Institute of Standards and Technology, Health and Environmental Sciences Institute, and Korean National Cancer Center.
Dr. Yovandich concluded by highlighting a recent visit with two patient families that have benefited from BDP therapies, which underscores the program's overall impact.

In the discussion, the following points were made:

- The BDP is pursuing an RNA nanoparticle–based therapy through the NExT Program; much of the technical work is being outsourced through Leidos Biomed. This work is in early stages, and the program is open to other collaborations in this space. FNLCR leadership has considered developing other types of platforms to enhance access to RNA-based technologies.

- GMP development and production is extremely expensive, and the science is generally well established for these clinical studies. Other partners, such as the members of the Can-ACT network, are well equipped to study scientific questions that can transition to a clinical trial.

- The BDP is interested in pursuing CAR natural killer–cell trials, but its activities are driven by current needs. The program's facilities have the capabilities for work in this area.

- The Can-ACT network is focused on centralized manufacturing; logistics establishment; and well-controlled, repeatable, robust manufacturing and analytical processes. The team has established a cryopreservation-based logistics process with the appropriate chain of custody documentation.

- A workshop or symposium highlighting FNLCR's activities in this space could provide valuable information to the research community.

- Access to BDP's services is controlled through associated networks and programs. Availability is strained by recent increases in demand.

IV. AI-DRIVEN MULTI-SCALE INVESTIGATION OF THE RAS/RAF ACTIVATION LIFECYCLE (ADMIRRAL) PROJECT—DRS. FRED STREITZ AND DWIGHT NISSLEY

Dr. Fred Streitz, Chief Computational Scientist, Lawrence Livermore National Laboratory (LLNL), and Dr. Dwight Nissley, Director, Cancer Research Technology Program, FNLCR, discussed the ADMIRRAL Project. They first provided an overview of the NCI–U.S. Department of Energy (DOE) collaboration, which started in 2016 and is in its second 5-year Memorandum of Understanding period among various national laboratories, including the FNLCR, LLNL, Argonne National Laboratory (ANL), Oak Ridge National Laboratory (ORNL), and Los Alamos National Laboratory. This collaboration has supported several projects, including ADMIRRAL, Innovative Methodologies and New Data for Predictive Oncology Model Evaluation (IMPROVE), Modeling Outcomes Using Surveillance Data and Scalable Artificial Intelligence for Cancer (MOSSAIC), and CANcer Distributed Learning Environment (CANDLE).

ADMIRRAL was started in collaboration with NCI and is a part of the RAS Initiative. The LLNL provides capabilities for high-performance computing at scale. The team is performing molecular dynamic (MD) simulations verified by experimental results. Dr. Nissley briefly highlighted the RAS/RAF activation life cycle. He explained that the primary effector of MAPK signaling is RAF kinase, which exists in an autoinhibited form; autoinhibition is released when the RAF kinase interacts with RAS, and subsequently the membrane, resulting in dephosphorylation, rearrangements of chaperone proteins, and structural changes, at the plasma membrane. Two RAFs then dimerize through kinase domains, resulting in a kinase-active signaling platform. The team believes that mechanistic insight to this process may present therapeutic opportunities to perturb oncogenic signaling.

The first stage of the project focused on building a model to better understand RAS in the context of a realistic membrane. The team performed simulations in the 8-lipid membrane at scale and identified fingerprints, or concentrations of specific lipids around RAS. These results were verified experimentally through various approaches. Drs. Streitz and Nissley explained, however, that several aspects of this process are challenging to model experimentally. Using the Multiscale Machine Learned Modeling Infrastructure (MuMMI), they performed hundreds of thousands of simulations utilizing millions of GPU-hrs on the largest computers in the world. The process is handled via specialized codes, which are open domain. Drs. Nissley and Streitz emphasized that ADMIRRAL is a highly collaborative, interactive, and multidisciplinary project that crosscuts multiple laboratories and entities. They concluded by commenting that this infrastructure could be applied for modeling other biological systems.

In the discussion, the following points were made:

- Potential protein rearrangement and domain movement is assessed in latent space using various models with the goal of determining the most appropriate trajectory and path from one state to another as defined through reaction coordinates.

- The team attempted to minimize the number of model parameters to control computational time. A dynamical density functional theory–based macro model has been published and is rigorously based on correlation functions that result from MD simulations.

- The macro model is driven by correlation coefficients from thousands of macro models. The team has developed a framework for executing a multiscale model on an exascale computer. These models could be carried out in various biological and physical contexts.

- Quantum computing could be considered in the future to accelerate the modeling process, and the LLNL is exploring developments in this area.

- The team defined an 8-lipid component membrane that was based on the prevalence of lipids in cells; these lipids are available experimentally and can be examined computationally. The outputs are being validated in cells and biochemically in defined experimental systems.

- Protein engineering is being used to modify residues at appropriate locations for attaching fluorescence resonance energy transfer components. The initial measurements for this work are in progress.

V. INNOVATIVE METHODOLOGIES AND NEW DATA FOR PREDICTIVE ONCOLOGY MODEL EVALUATION (IMPROVE)—DRS. RICK STEVENS AND M. RYAN WEIL

Dr. Rick Stevens, Associate Laboratory Director for the Computing, Environment and Life Sciences, ANL, and Dr. M. Ryan Weil, Director, Strategic and Data Science Initiatives, FNLCR, presented on the IMPROVE project, which is focused on analyzing and improving predictive models for tumor drug response. They explained that AI offers potential for predicting the outcome of cancer therapies, particularly chemotherapy, but this technology is still in early development. IMPROVE is focused on trustworthiness and reproducibility of the AI being used in this application.

NCI and DOE offer unique capabilities in this effort: DOE has assembled large teams working on machine-learning and AI methods, and NCI has assembled teams that are familiar with cancer, experimental cancer data, and treatment strategies. The outcome of these efforts is a standardized framework that allows researchers to study and compare models for tumor drug response and other applications. IMPROVE has two aims: (1) Build a system for curating models being developed by the research community and develop capabilities able to compare the models using supercomputers, and (2) work in partnership with the community to develop protocols for specifying drug-screening experiments and improving models, and apply those protocols to generate new data aimed at improving model performance.

Dr. Stevens outlined the project workflows, which are focused on model curation, data curation, and framework development. He highlighted examples of the analysis, which depict model generalization. Understanding the aspect of the model that is contributing to this variability in generalization is the key insight, and models that show in-study generalization often are not the best models for out-of-study generalization; thus, groups that are evaluating or producing a new model need to train their model and test it on different data sets. The team is applying this approach for patient-derived organoids and patient-derived xenografts.

Hyperparameters structure the models (e.g., learning rates) and can provide additional insights for researchers. Most groups are not performing hyperparameter optimization currently. The team has been developing CoderData, a data package that is an aggregation of essentially all the types of data that the community is using to build tumor drug responses. These efforts are making community data more AI ready and reducing the data management burden. Community participants have been engaged in curating the data, which will be made publicly available.

Versions of the IMPROVE framework are ready for use, and the ecosystem is growing rapidly through collaborative efforts. The team is prioritizing strategic collaborations to amplify investments and reduce duplicate work. These relationships are bidirectional, and the response data generated by IMPROVE are benefiting the programs and research community. Additionally, the team is helping drive

their collaborators' research experiments and provides insights into AI applications. IMPROVE is interested in fundamentally understanding AI models, and collaborators can provide insights in this area.

In the spirit of amplifying existing efforts and furthering the AI readiness of data, IMPROVE is collaborating with a number of laboratories—including at the University of California, Los Angeles; University of California, San Diego; MD Anderson Cancer Center; Dana-Farber Cancer Institute; and The University of Chicago—that have collections of either large or unique cancer data. By working with these partners and performing cross-validation studies, the team can create links among the data for comparisons and future use. Currently, IMPROVE is performing initial pilot studies focused on colorectal cancer; initial studies have indicated that this cancer is poorly predicted by existing AI models. The team plans to expand to other types of cancer—including adenocarcinoma of the pancreas, glioblastoma multiform, AML, and sarcoma—as well as a broader range of drugs.

In closing, Drs. Stevens and Weil reiterated that from its beginning, IMPROVE was conceived to be outward facing; the resources being generated were intended to support community needs. The project involves multiple partners from various institutions and holds regular meetings and events to foster community engagement. All resources are open source, and the project website provides additional information for the community.

In the discussion, the following points were made:

- Existing biases (e.g., ethnic, racial, gender) are likely present in the IMPROVE data. The team is developing a tool to further explore these dynamics. AI tends to optimize for the most abundant classes, so minority classes often are excluded from the analyses. Model adjustments could help compensate for class imbalance, but more work in this space is needed.

- IMPROVE could incorporate patient-derived xenograft treatment response data from NCI's pediatric preclinical *in vivo* testing program. The program is moving away from traditional cell lines and towards more patient experience relevant model systems.

- Explainability is sometimes a challenge but has been addressed in recent years. Researchers can now build models that help explain the decision-making of the primary model, as well as the features that the model is using. Sophisticated features are available for normalizing and resampling the data, but this is still a challenge. The team is exploring how data augmentation and synthetic data can be used to mitigate some of these effects.

VI. MOSSAIC: ACHIEVING NEAR REAL–TIME CANCER SURVEILLANCE WITH AUTOMATIC RECORD ABSTRACTION—DRS. HEIDI HANSON AND BETSY HSU

Dr. Heidi Hanson, Group Leader, Biostatistics and Biomedical Informatics Group, ORNL, and Dr. Betsy Hsu, Chief, Surveillance Informatics Branch, Surveillance Research Program, Division of Cancer Control and Population Sciences (DCCPS), NCI, discussed MOSSAIC's impact in achieving near real–time cancer surveillance with automatic record abstraction. They explained that NCI's Surveillance, Epidemiology, and End Results (SEER) Program provides the framework for MOSSAIC's work. SEER was authorized by the National Cancer Act with the mission to support research, and the SEER registries submit deidentified data to NCI, which are made available to researchers through appropriate authentication and authorization processes.

SEER currently consists of 18 registries, and about 85 percent of the incident cases received by the SEER registries have associated electronic pathology reporting, which is critical to achieve near real–time incidence reporting. A cancer abstract in a registry is consolidated from various records, such as

hospital abstracts, physician reports, pathology reports, and death certificates. This information is becoming more complex and heterogeneous and includes new data sources (e.g., pharmacies, claims, genomic testing results, real-time data feeds). Overall, the process for screening and abstraction is largely manual. The SEER*Data Management System (SEER*DMS), the information technology infrastructure used by the registries, enables centralized data linkages to expand the breadth and depth of information collected on cancer patients, as well as simultaneous implementation of new tools—such as those that are being developed by MOSSAIC—allowing the team to support a common and optimized production workflow for the registries.

MOSSAIC is a key component to enhancing the SEER infrastructure to support a broader set of cancer research activities. The activities on MOSSAIC are centered around the development of AI workflows for extracting information in an automated fashion for cancer surveillance. This work supports near real–time cancer incidence reporting, as well as extracting information that will support better understanding of patient outcomes at the population level. With the growing complexity of cancer diagnosis and treatment, capturing information essential to understanding differences in outcomes in cancer patients—such as subsequent treatment, disease progression, or metastasis—is increasingly difficult. MOSSAIC is addressing this issue via near real–time incidence reporting through the development of AI tools that automatically abstract of tumor characteristics from pathology reports, near real–time case ascertainment through identification of cancer-related reports, and a better understanding of outcomes through prediction of metastasis, all of which are important in capturing essential information about patients.

New methods developed by MOSSAIC represent opportunities to close critical information gaps and enable rapid phenotyping of cancer data to improve patient outcomes. The AI tools they have developed (BARDI and FrESCO) are being expanded to include other critical information for health outcomes, such as exposomic measures. One of the initial focus areas of MOSSAIC was to develop deep learning models to automatically extract tumor characteristic information from pathology reports, as this information is critical for incidence reporting. Using these deep learning models, the team is currently autocoding about 29 percent of pathology reports at a 98 percent accuracy threshold, representing a time savings of nearly 14,000 person-hours per year. This pathology extraction application programming interface (API) is currently used in production across 15 SEER and four non-SEER registries, with up to three new registries expected in the next 2 years. Registries have also found uses of the results from the API to increase operational efficiencies in ways that were not anticipated initially (e.g., rapid case ascertainment, research studies, staff training).

The MOSSAIC partnership has been synergistic and brings a large volume of high-quality data to a combination of subject-matter experts, high-performance computing experts, and software engineers to develop innovative solutions for near real–time disease surveillance in the cancer community. Over the past two years, the team has broadened its focus, aiming to develop tools that can be used by the broader oncology community; they are coupling innovative research under the development of modular software design—an approach that structures software as a collection of distinct components, or modules, that are designed to be interchangeable for maximum flexibility and reusability in other contexts. This enhanced flexibility and adaptability accelerates the research and development process, leading to faster breakthroughs and quicker implementation of research outcomes.

Dr. Hanson provided examples of work using current production algorithms that are operating in the SEER registries. She described Batch-processing Abstraction for Raw Data Integration (BARDI), the program's tool for AI readiness for clinical data. The team uses BARDI to transform unstructured pathology report data used by MOSSAIC into AI ready features through common data preprocessing operations. BARDI simplifies both the development and upkeep of complex data pipelines, making it efficient to use for both MOSSAIC and other projects. Overall, this allows the team to develop a

systematic, flexible, and reproducible pipeline for making data AI ready. Additionally, the Framework for Exploring Scalable Computational Oncology (FrESCO) takes the data that have been tokenized and can be used to train a range of deep learning prediction algorithms for several different tasks, including reportability, site, subsite, laterality, behavior, recurrence, metastasis, biomarkers, and treatment. This package has been used for retraining models, improving their accuracy in predicting reportable reports. Projects in the Department of Veterans Affairs, British Columbia Cancer Registry, and Kentucky Cancer Registry have used BARDI and FrESCO to create their own models for cancer classification.

The team has developed the Path-BigBird model, a large-language transformer model that was trained from scratch using pathology reports from SEER registries. The language in pathology reports differs from the language generally used in scientific reporting; having a foundational model trained specifically in that type of language yields better prediction accuracy than using publicly available large-language models, particularly on tasks that are challenging to predict (e.g., histology). The team also is developing a phrase-level attention module that can be coupled with foundational models for oncology to improve predictions. The phrase level attention module differs from traditional word or token level attention because it focuses on the phrases; allowing it to be more aware of the contextual information. In addition, the team is developing a new mixture of expert models that ensemble the predictions from multiple models to achieve higher accuracy. This framework can be used for heterogenous ensembling of models as well. Dr. Hanson emphasized that overall, these promising scientific developments will help the team achieve an autocoding rate of 50 percent in the near future.

Identification of pathology reports related to cancer is another critical piece to achieving near real–time incidence reporting because cancer registries are authorized to hold information only about cancer cases. MOSSAIC is helping to achieve this through development of the reportability API. Currently, the team offers multiple algorithms that can be used to filter cancer-related reports from facilities before they are reported to registries; this new solution improves upon existing methods that can often lead to a large number of false positives that must still be manually reviewed by the registries to confirm a reportable case. Improving the ability to automatically screen for cancer-related reports is essential to the team's goals of rapid case ascertainment for research studies, as well as for achieving near real–time incidence reporting. Currently, they are validating the performance of this reportability API across a broader set of registries.

In regard to near real–time cancer case identification, the team was able to develop a model using information from nearly 4 million cancer pathology reports. This algorithm had high accuracy in identifying reportable and nonreportable cases. The team also has developed algorithms for near real–time recurrence and metastasis. Currently, no standard exists for population-level collection of recurrence in metastatic disease, which makes it difficult to assess risk and follow patients longitudinally. The MOSSAIC team developed an algorithm that classifies pathology reports for both recurrence and metastasis. They collected information from more than 67,000 reports, allowing them to train a highly accurate model for both recurrence and metastasis that is currently being tested for deployment at the SEER registries.

The team has been developing real-world privacy preserving measures for federated learning developments. Federated learning serves as a solution to data siloing that exists within health care research, but major scientific challenges still exist (e.g., the trade-off between privacy and accuracy). SEER data can be used for testing real-world solutions in a translational way. A privacy-preserving hackathon will be held in June 2024, with a focus on measuring privacy leakage when training large language models with federated learning. Privacy leakages may result in in the reidentification of patients. Other current activities are focused on linking the external exposome to residential histories of cancer patients. The team will continue to combine information across many different social and environmental measures and link these data to over 30 years of residential history data. They are creating a database that

spatially links multiple social and environmental determinants of health datasets and can be utilized to assign an exposure profile to a cancer patient prior to and after cancer diagnosis; improving models of cancer incidence and survival. This is a broader effort that involves multiple partners including the VA and DOE.

Essentially, MOSSAIC aims to develop real-world AI solutions for population surveillance. These solutions are applicable to both siloed and centralized worlds and can accelerate identification of disease occurrence to population-level disease statistics. This effort is important for population health surveillance and can yield real-world solutions for precision health. Clinicians can link specific risk factors to individuals and examine how those factors affect prognosis and trajectories. They can then apply those solutions for population health surveillance, as well as in the clinic. All of these solutions are publicly available.

In the discussion, the following points were made:

- SEER has been working to bring in additional linkages through partnerships with outside data vendors. As SEER expands the kind of information that is available, it also provides additional data that can be used for modeling through MOSSAIC. The team is designing the system so that it can be applicable outside of SEER data. Additionally, the FrESCO framework is being used as a foundation for work for bio-surveillance because these models can be used across different modalities.

- Variation in nomenclature can be a challenge. The team does not perform rule-based mapping; they have set rules that clean the data, and the data are preprocessed and tokenized by word pieces or subwords. They are building on other tokenization modules that allow them to connect to broader space within the transformer AI world. The team is building an AI toolbox with interoperable modules. Key to this toolbox is the BARDI module, which takes raw unstructured clinical text and transforms it to AI ready features. FrESCO contains modules that allow for model training with a convolutional neural networks, deep neural networks, BERT based transformer models (foundational models), and federated learning. This serves as a foundational and flexible ecosystem for oncologic data classification. The team has examined recurrence predictions via a multimodal recurrence model. The models also have a built-in abstention. Recurrent metastatic disease is an easier endpoint to identify, and label noise has posed a challenge for scoring.

- The registries can hold any information about a cancer patient, including follow-up pathology reports, and pathology laboratory facilities process pathology reports for many diseases in addition to cancer; registries maintain those reports. Many cancer registries are adjusting their policies so that they can receive information on non-cancer-related pathology reports, which can be used as part of the screening process. Diagnoses potentially could be shared across registries. Additionally, registries can contain prediagnosis patient history, which could be linked to exposure data. Some registries capture information on premalignant lesions, but these collections are not consistent across registries.

VII. CLOSING REMARKS—DR. CANDACE S. JOHNSON

Dr. Johnson expressed appreciation to the Committee members and other participants for attending. Members were reminded to send potential agenda topics for future FNLAC meetings to Dr. Kane.

VIII. ADJOURNMENT—DR. CANDACE S. JOHNSON

There being no further business, the 15th Virtual Meeting of the FNLAC was adjourned at 4:11 p.m. EDT on Monday, 11 March 2024.

7/11/2024 /s/

Date Candace S. Johnson, Ph.D., Chair

7/11/2024 /s/

Date Christopher D. Kane, Ph.D., Executive Secretary