Data Science Initiatives at the NCI

Norman E. Sharpless, M.D.
Director

Computational Approaches for Cancer Workshop
November 13, 2020

@NCIDirector
@TheNCI
Today

• NCI Updates
• COVID-19 Research at NCI
• Select NCI Data Science Efforts
• A Look Ahead
NCI Appropriations
FY 2015 – 2020 (in millions)

21st Century Cures Act - orange
Childhood Cancer Initiative - green


$4,950 | $5,215 | $5,689 | $5,965 | $6,144 | $6,440 | $6,249

$0 | $1,000 | $2,000 | $3,000 | $4,000 | $5,000 | $6,000

+ $306M COVID-19 serology (April 2020)
+ $414M emergency funding

House Mark

$6,195 | $6,440 | $6,494

$195 | $195 | $195

$50 | $50 | $50
NCI Research Project Grants (RPG) Funding and R01 Paylines

- RPG funding levels exclude small business grant set-asides.
- FY 2021 appropriations not yet finalized.
NCI Response to COVID-19

- SARS-CoV-2 serology research
- NCI COVID-19 in Cancer Patients Study (NCCAPS)
- Guidance and special procedures for cancer clinical trials
- Flexibilities and opportunities for grantees
- Genomic studies of COVID-19 outcomes

[cancer.gov/coronavirus-researchers]
Why NCI?
SeroHub: SARS-CoV-2 Serology Study Dashboard

- Effort began in early June at the request of HHS, CDC, and NIAID to develop a data warehouse & dashboard for tracking SARS-CoV-2 seroprevalence and other US-based serology studies
- Builds on FNL expertise in warehouse and dashboard development for other NCI resources

Key features

- Publicly accessible data warehouse to systematically document and track SARS-CoV-2 serology studies and associated test results
- Tracking dashboard to visualize SARS-Cov-2 serology data and present results overall and by key strata
SeroNet
Serological Sciences Network

- Grants (U54s and U01s)
- Contracts (CBCs)
- National Laboratory

Launched October 8
Geographical Distribution of SeroNet Sites

- U54 Site
- U01 Site
- CBC Site
COVID-19 and cancer

With the spread of coronavirus disease 2019 (COVID-19), societies and states have in many instances described our efforts to stop the disease and its spread as the situation in the United States, including people from screening, diagnosis, and treatment for non-COVID-19 diseases. The management of cancer patients, for example, could be substantially different. What can be done to minimize this effect?

Cancer is a complex set of diseases whose prognosis varies with the type of cancer and its stage of development. In general, the earlier one receives cancer treatment, the better the outcome. However, there have been a number of developments in cancer diagnosis in the United States since the start of the pandemic, including a drop in numbers of new cancer diagnoses and treatments. This drop in new cancer diagnoses and treatments has also led to a delay in cancer treatment for some patients. The situation is worsening. At many hospitals, cancer surgeries and biopsies are being canceled due to the potential for COVID-19 exposure. This has led to an increase in the number of patients with advanced cancer diagnoses and a delay in cancer treatments. The situation is worsening. At many hospitals, cancer surgeries and biopsies are being canceled due to the potential for COVID-19 exposure. This has led to an increase in the number of patients with advanced cancer diagnoses and a delay in cancer treatments.

NORMAN E. SHAPIRO

**Modeling and simulating excess deaths from colorectal and breast cancers, 2020 to 2030**

Colorectal Breast

<table>
<thead>
<tr>
<th>Year</th>
<th>Colorectal</th>
<th>Breast</th>
</tr>
</thead>
<tbody>
<tr>
<td>2020</td>
<td>6000</td>
<td>3000</td>
</tr>
<tr>
<td>2025</td>
<td>5000</td>
<td>2500</td>
</tr>
<tr>
<td>2030</td>
<td>4000</td>
<td>2000</td>
</tr>
</tbody>
</table>

By Laura McDermott

**The Washington Post**

By Laura McDermott

**STAT**

Ignoring cancer care now may trade one public health crisis — Covid-19 — for another, NCI chief warns

By Elizabeth Cooney
Supporting Cancer Research during the Pandemic

Impacts on cancer care delivery and outcomes

Studying COVID-19 in people with cancer

Adjusting clinical trial protocols

Cushioning the blow for cancer researchers

Outlook for research funding beyond the pandemic
Flexibilities to support grantees during the pandemic

- Extending deadlines for applications
- Allowing institutions to use NCI grant funds to maintain salaries and stipends
- Extending project timelines and reporting requirements
- Extending eligibility periods for early-stage investigators and trainees
- Carryover for institutional training grants (T35, T32, K12) with prior approval
Key Focus Areas

**BASIC SCIENCE**
Reaffirm our commitment to basic science to drive novel approaches and technologies

**WORKFORCE DEVELOPMENT**
Support the cancer research enterprise by focusing on the workforce of cancer investigators

**BIG DATA**
Increase data aggregation and interpretation to speed our work across the cancer enterprise

**CLINICAL TRIALS**
Fully realize the power of clinical trials through innovative design, administration, and analyses
Real World Evidence

HPV Vaccination and the Risk of Invasive Cervical Cancer

Jiayao Lei, Ph.D., Alexander Ploner, Ph.D., K. Miriam Elfström, Ph.D., Jiangrong Wang, Ph.D., et al.

October 1, 2020
NCI Cancer Research Data Commons

Cloud-based data science infrastructure that provides secure access to a large, comprehensive, and expanding collection of cancer research data.
Datasets in the CRDC

And many more…
Clinical Trials Data
NCTN/NCORP Data Archive - Overview

A centralized, controlled-access database that contains individual, patient-level data for all phase 3 & phase 2/3 clinical trials with:

- Primary Results publications available on or after January 1, 2015 and selected non-primary publications available on or after April 1, 2018
- Other data (e.g., phase 2 trials or legacy phase 2 & phase 3 trials published before 2015, etc.) are included on a case-by-case basis.
- Submissions are due within 6 months of publication date with all data elements presented in the publication included
- All data available within the Archive are also accessible via Project Data Sphere (PDS) within 24 hours

Complements other NCI data sharing activities that focus on genomic or imaging data as well as sharing of biospecimens from NCTN trials for approved correlative science proposals (i.e., Navigator).
Current NCTN Navigator Biospecimen Inventory (Oct 2020)
from completed Phase 3 trials; available for approved research proposals

![Pie chart showing distribution of biospecimens by age and type]

- **Pediatric - ALL**: 22%
- **Pediatric – Other Heme**: 3%
- **Pediatric – Other**: 8%
- **Adult - Breast**: 27%
- **Adult - Colorectal**: 7%
- **Adult - Lung**: 3%
- **Adult - Ovarian**: 8%
- **Adult - Prostate**: 7%
- **Adult - Hematologic**: 3%
- **Adult - Other**: 12%

**Total Specimens**: 199,921
**Patients**: 146,562
**Trials**: 231

[navigator.ctsu.org](http://navigator.ctsu.org)
NCTN/NCORP Data Archive – **New Data Integration**

- For RTOG-0617*, a data integration is in place with The Cancer Imaging Archive, NCI's official imaging repository for NCTN trials, that enables users to request both clinical and imaging data.

- Images from 3 additional trials are being de-identified.

*High-Dose or Standard-Dose Radiation Therapy and Chemotherapy With or Without Cetuximab in Treating Patients With Newly Diagnosed Stage III Non-Small Cell Lung Cancer That Cannot Be Removed by Surgery*
Molecular Profiling to Predict Response and Treatment (MP2PRT)

Challenge
Gathering information from clinical trials to inform new trial design takes too long, particularly in rare cancer types or in specific populations.

Through retrospective analysis of archival specimens collected through clinical trials, develop predictive models that will identify what patients might benefit from standard therapy or who might require additional or novel interventions.
# Molecular Profiling to Predict Response and Treatment (MP2PRT)

<table>
<thead>
<tr>
<th>Group</th>
<th>Disease</th>
<th>Description</th>
<th>Investigator</th>
</tr>
</thead>
<tbody>
<tr>
<td>Children’s Oncology Group (COG)</td>
<td>Acute Lymphoblastic Leukemia (ALL)</td>
<td>Comprehensive Genomic Profiling to Identify Alterations Associated with Relapse for NCI Standard-Risk B-lineage ALL and NCI High-Risk B-lineage ALL with Favorable Genetic Features</td>
<td>Mignon L. Loh, M.D. Benioff Children’s Hospital, University of California SF, CA</td>
</tr>
<tr>
<td>Children’s Oncology Group (COG)</td>
<td>Wilms Tumor</td>
<td>Identification of Genetic Changes Associated with Relapse and/or Adaptive Resistance in Patients Registered with Favorable Histology Wilms Tumor on AREN03B2 (Renal Tumors Classification, Biology, and Banking Study)</td>
<td>Elizabeth Perlman, M.D. Anne &amp; Robert H. Lurie Children’s Hospital of Chicago, IL</td>
</tr>
<tr>
<td>NRG Oncology Group</td>
<td>Cervical Cancer</td>
<td>Genomic and molecular characterization of biomarkers associated with tumor angiogenesis, DNA repair, and immunologic tolerance among exceptional responders and long-term survivors in NRG /GOG protocol 240</td>
<td>Krishnasu Tewari, M.D. UC Irvine Orange, CA</td>
</tr>
<tr>
<td>NRG Oncology Group</td>
<td>Meningioma</td>
<td>Identifying novel molecular markers of response to radiotherapy in meningiomas using samples from the RTOG-0539</td>
<td>Kenneth Aldape, M.D. Center for Cancer Research, NCI</td>
</tr>
</tbody>
</table>

[NIH NATIONAL CANCER INSTITUTE]
NCI Cloud Resources

- The Institute for Systems Biology, Seven Bridges Genomics, and Broad Institute are referred to as NCI Cloud Resources (CRs)
- 3 CRs connect NCI data and compute in the cloud
- Access to data, workspaces, analysis tools, and pipelines
- Ability for researchers to bring their own data and tools
- Steady increase in active community of cancer researchers adopting the Cloud Resources for their data analysis
  - **Average of 3000 active users per month across the CRs**
  - **4 Million workflows run during 2020**

- NCI Cloud Resources
  - 23+ Datasets
  - Elastic Compute
  - FISMA moderate
Cancer Data Aggregator

Enables researchers to discover, query, retrieve, and aggregate data across CRDC and NCI DCCs data according to a variety of search parameters.
Surveillance, Epidemiology, and End Results (SEER) Program

- Recent RFP to expand the SEER program further (contract January 2021)
- With new registries—550,000 incident cases received annually
  - Approximately 85% of cases with real time electronic pathology (e-path) reporting
  - Facilitates rapid case identification supporting research
- All registries will be on a common data platform (SEER DMS) that permits
  - central linkages with external partners
  - facilitates scaling of new initiatives across all registries simultaneously

16 population-based registries now covering 35% of the US population
Partnerships and linkages to enhance SEER data

Partnering to acquire source data

Genomic/genetic testing
GHI, Caris LS, Decipher, Castle Life Sciences, Myriad, Ambry, etc.

Claims data

Unlimited Systems - oncology claims processor ~15% of cases in SEER
Large insurers (United Health Care)
Exploring All Payer All Claims (6 SEER registries have state-wide APAC)
Pharmacy (CVS and Walgreens/RiteAid)
  Data received for all registries- 2014+

Partnerships with technology companies aggregating and using clinical data

- CancerLinQ, Syapse, Tempus (provider EMR data)
- Varian/Elekta (radiation oncology)
- Ambra Health (radiology)
During the initial years (2010-2012), there was some evidence of differential testing by race and ethnicity. Recent data suggests disparities are disappearing.
CCDI
BUILDING A COMMUNITY CENTERED AROUND CHILDHOOD CANCER CARE AND RESEARCH DATA

LEARN FROM EVERY CHILD

DATA TYPES:

CLINICAL
TREATMENT
OUTCOME
MOLECULAR
BIOSPECIMEN
LONGITUDINAL
POPULATION

PROGRAM GOALS:

• DEVELOP NEW TREATMENTS
• IMPROVE EXISTING TREATMENTS
• IMPROVE OUTCOMES, QUALITY OF LIFE, AND SURVIVORSHIP

IMPROVING THE QUALITY, CONSISTENCY, ACCESS, AND ACCESSIBILITY OF DATA FROM EVERY CHILD WILL ALLOW US TO ACHIEVE OUR GOALS.
# National Childhood Cancer Registry

## Initial Registry Participation (70% of US childhood cancers)

### 7 NPCR Registries
- Florida
- Illinois
- New Jersey
- Ohio
- Pennsylvania
- Tennessee
- Texas

- 2020 - submission of de-identified NAACCR data in NCCR
- 2021 - full registry submission with PII to DMS*Lite as a repository for linkages and submission to the NCCR

### SEER Registries
- Georgia
- Los Angeles
- Greater CA
- Greater Bay Area
- Iowa
- Connecticut
- Kentucky
- Louisiana
- Seattle
- Idaho
- New York
- Massachusetts

- SEER will contractually require submission to the NCCR

### Goal
- 100% coverage of all pediatric patients over the next few years
NCI-DOE Collaboration: Joint Design of Advanced Computing Solutions for Cancer (JDACS4C)

DOE-NCI partnership to advance exascale development through cancer research
## NCI-DOE Collaboration (JDACS4C)

<table>
<thead>
<tr>
<th>PILOT 1</th>
<th>CANDLE (CANcer Distributed Learning Environment)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predictive Modeling for Pre-Clinical Screening</td>
<td>An exascale computing project to develop machine learning framework for cancer</td>
</tr>
<tr>
<td>Develop reliable machine-learning based predictive models of anti-cancer drug response</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>PILOT 2</th>
<th>ATOM (Accelerating Therapeutics for Opportunities in Medicine)</th>
</tr>
</thead>
<tbody>
<tr>
<td>RAS Biology in Membranes</td>
<td>Developing a pre-clinical drug design and optimization platform that leads with computation to help shorten the drug discovery timeline</td>
</tr>
<tr>
<td>Develop multiscale modeling capabilities to investigate RAS dynamics on cell membranes</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>PILOT 3</th>
<th>UNCERTAINTY QUANTIFICATION</th>
</tr>
</thead>
<tbody>
<tr>
<td>Population Information Integration, Analysis, and Modeling Information; capture of unstructured clinical text using Natural Language Processing (NLP) and Deep Learning algorithms</td>
<td></td>
</tr>
<tr>
<td>General methods to improve confidence/level of certainty of results from predictive computational models from JDACS4C pilots</td>
<td></td>
</tr>
</tbody>
</table>
MD simulation of KRAS4b suggests the hypervariable region (HVR) forms binding pocket with the G domain.
What Lies Ahead? Unresolved questions in cancer research where high-performance computing may be critical

**Structural Biology**

predicting T cell receptor binding to MHC + antigen, predicting Ab structure

**Medicinal Chemistry and Drug Discovery**

in silico docking studies, predicting in vivo toxicity, etc.

**Novel Analytics of Large Complex Datasets**

proteomic-genomic-clinical data; large deidentified sets of clinical and personal data linked via unique identifier technologies

**Modeling of complex, multi-parameter systems within a large search space**

models of hematopoiesis, drug resistance in population of inter-dependent tumor cells, etc.
Discussion