FNLAC NCI-DOE Collaborations *ad hoc* Working Group Report

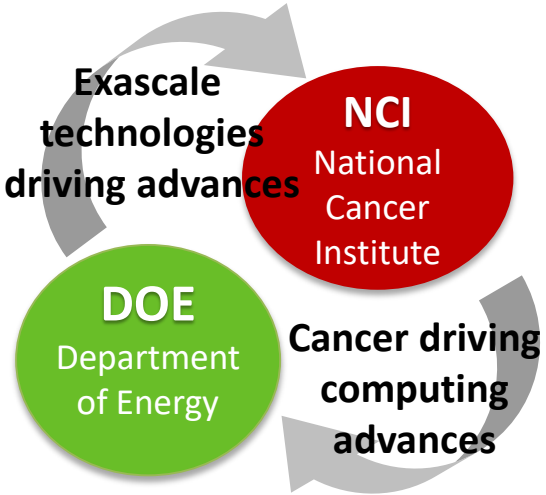
Dr. Piermaria J. Oddone for the Working Group

Meeting of the Frederick National Laboratory Advisory Committee
June 27, 2019

NCI-DOE Collaboration

The NCI/DOE Collaborations were formed to jointly accelerate NCI and DOE federal missions in precision oncology and high-performance computing (HPC).

The partnership is designed to push the frontiers of high performance computing through application to NCI's mission to improve understanding of cancer biology and its application to more effective cancer therapies.



NIH NATIONAL CANCER INSTITUTE

Argonne NATIONAL LABORATORY

OAK RIDGE

Lawrence Livermore National Laboratory

Los Alamos NATIONAL LABORATORY EST. 1943

Frederick National Lab for Cancer Research

Mission of the Working Group

- Provide scientific evaluation of programs, projects and activities formed in support of or relevant to NCI-DOE collaborations
  - Specifically provide suggestions to the three pilot projects on how to optimize their impact for both DOE and NCI programs.
  - Provide guidance and insights on relevant partnerships with other entities (e.g. ATOM)
  - Explore the new domains and activities in which collaborations between the NCI and DOE would be mutually beneficial and advance the missions of these entities

- The Working Group will advise the FNLAC
  - In accordance with the NCI/DOE MOU, the DOE Secretary, DOE and DOE FACA committees may use the public products and public findings in furthering the DOE mission

Activities under the DOE-NCI Collaboration

Joint Design of Advanced of Computing Solutions for Cancer (JDACS4C)

Uncertainty Quantification

- **Cellular Level Pilot 1:** Predictive Models for Pre-clinical Screening
- **Molecular Level Pilot 2:** RAS Biology in Membranes
- **Population Level Pilot 3:** Population Information Integration, Analysis, and Modeling
- **CANDLE** (CANcer Distributed Learning Environment)**:** An Exascale Computing Project to develop Machine Learning framework for Cancer

Accelerating Therapeutics for Opportunities in Medicine (ATOM)

Today's update

Joint Design of Advanced of Computing Solutions for Cancer (JDACS4C)

Uncertainty Quantification

- **Cellular Level Pilot 1:** Predictive Models for Pre-clinical Screening
- **Molecular Level Pilot 2:** RAS Biology in Membranes
- **Population Level Pilot 3:** Population Information Integration, Analysis, and Modeling
- **CANDLE** (CANcer Distributed Learning Environment)**:** An Exascale Computing Project to develop Machine Learning framework for Cancer

Accelerating Therapeutics for Opportunities in Medicine (ATOM)

Today's update

- An update on the three pilots and associated uncertainty quantification (the last update to FNLAC was a little over a year ago):
  - Accomplishments and lessons learned in years 1-3
  - The WG views on the aims and direction for years 4-5

- **Two notes**:
  - Each pilot is really a major collaborative undertaking, exploring a very difficult problem, that will require a sustained effort over many years. Calling them *major projects* would be just as appropriate as calling them pilots.
  - Continuation of projects will aid the broader NCI community in deriving impact and value from the partnership.

Pilot 1

Predictive Models for Pre-Clinical Screening

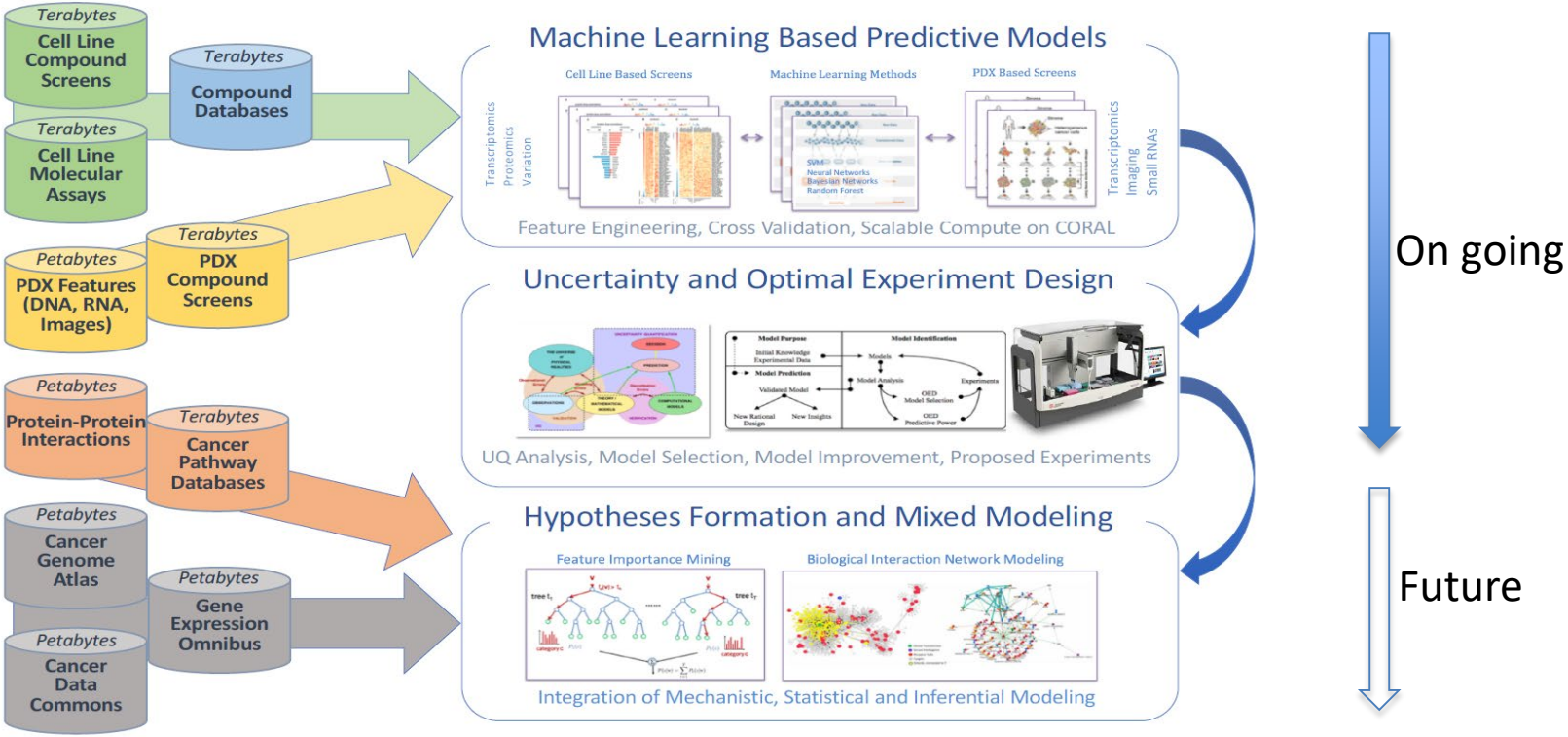
Described a year ago:
Aim 1: Develop reliable machine-learning-based predictive models of anti-cancer drug response
Aim 2: Integrate uncertainty quantification and optimal experimental design to assert quantitative limits on predictions
Aim 3: Develop hybrid predictive models that support the graded introduction of mechanistic models into the machine-learning framework

Within the global aims described a year ago, specific aims for year 4 and 5 are:

1. Advance state-of-the-art machine learning models for PDX and PDO drug response predictions
2. Develop low data learning methods aimed at maximizing the value of high-cost experiments
3. Develop methods for Interpretability of deep learning models for hypothesis formation and explainability

Pilot 1: Predictive models for pre-clinical screening

Foundation has been established with large, complex and accessible system



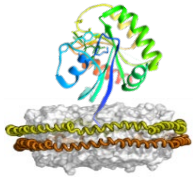
Pilot 1

The narrowing and focusing of goals for year 4 and 5 follows from accomplishments and lessons learned by the collaboration in years 1-3 and is consistent with the advice of the working group.

1. The ultimate limited utility of CL based predictions means moving to PDO and PDX models
2. Limited availability of PDO and PDX models means understanding ML and UQ in the context of limited statistical samples
3. Main problem is not algorithmic but data. Formalize definition of appropriate (more/better) data for ML (minimum requirements) – describe datasets that would be most useful in language that is informative to cancer biologists
4. Review transfer learning from CL to PDO to PDX. How useful?

Pilot 2

RAS Biology in Membranes

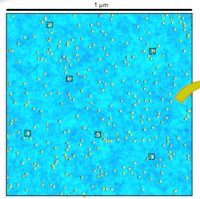


Described a year ago:
Aim 1: Develop multiscale modeling capabilities to investigate RAS dynamics on cell membranes
Aim 2: Understand how RAS and extended RAS complexes are activated and simulate RAS-RAF interactions on realistic, lipid-bilayer membranes
Aim 3: Develop machine learning-enabled dynamic model validation approach to high-fidelity simulation

Extended aims for years 4 and 5:

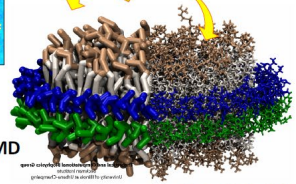
1. Spatial hierarchical multi-scale modeling
   - Extend the two-scale simulation approach to a third scale: atomistic resolution
   - Extend the macroscopic model to allow incorporation of membrane curvature
2. Understand activation of extended RAS complex
   - Deeper understanding of the RAS signaling cascade and RAF-RAS interactions
   - Extend workflow to enable simultaneous communication across the three scales
3. Machine-learning enabled dynamic validation approach to high-fidelity simulation
   - Define ML approach to handle scale transition decisions across three levels

Pilot 2: Ras Biology in Membranes



Macro model

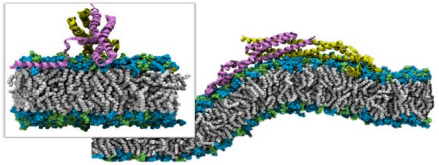
Three-levels of multiscale



CGMD

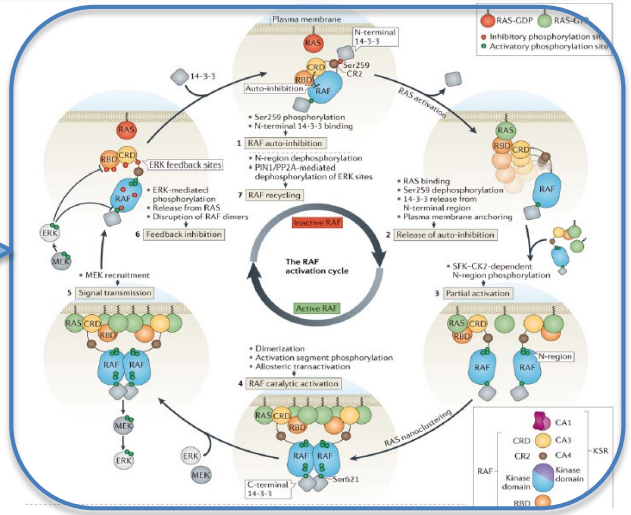
Atomistic MD

Realistic membrane topology



https://www.psc.edu/science/2007/bardomain/

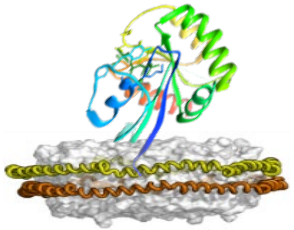
Curvature is implicated in RAS biology,



Now: 2 levels

Future: 3 levels

Future: deeper understanding of RAS signaling cascade

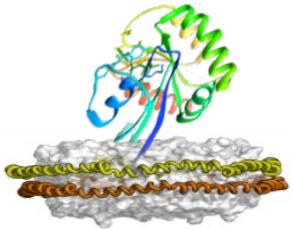


Pilot 2

The aims of Pilot 2 are largely unchanged. Progression to an atomistic level is necessary to start understanding in more detail how RAS4B binds to other molecules on the lipid surface, thus starting the signaling cascade.

Experiments on artificial membranes provide data to verify the computational models and structural data on various proteins allows the modeling of those interactions at the atomistic level.

The computational model pinned in selected cases with these actual experiments is filling the gaps that cannot be seen experimentally in the interactions of KRAS4B with the lipid membrane and other proteins.



Pilot 2

The Working Group is impressed with the progress in the two scale simulations that have given some insight into the behavior of KRAS4B on the membrane

Going to three scales by adding an atomistic simulation should greatly expand capabilities. It is important that the plans be quite detailed and clear for the exchange from biology to simulations and back – filling gaps in experimental design and vise versa.

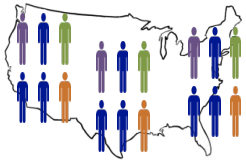
Ab-initio modeling of such a large experimental system will require sustained effort to get to the point of learning new biology and new potential therapeutic targets

Pilot 3

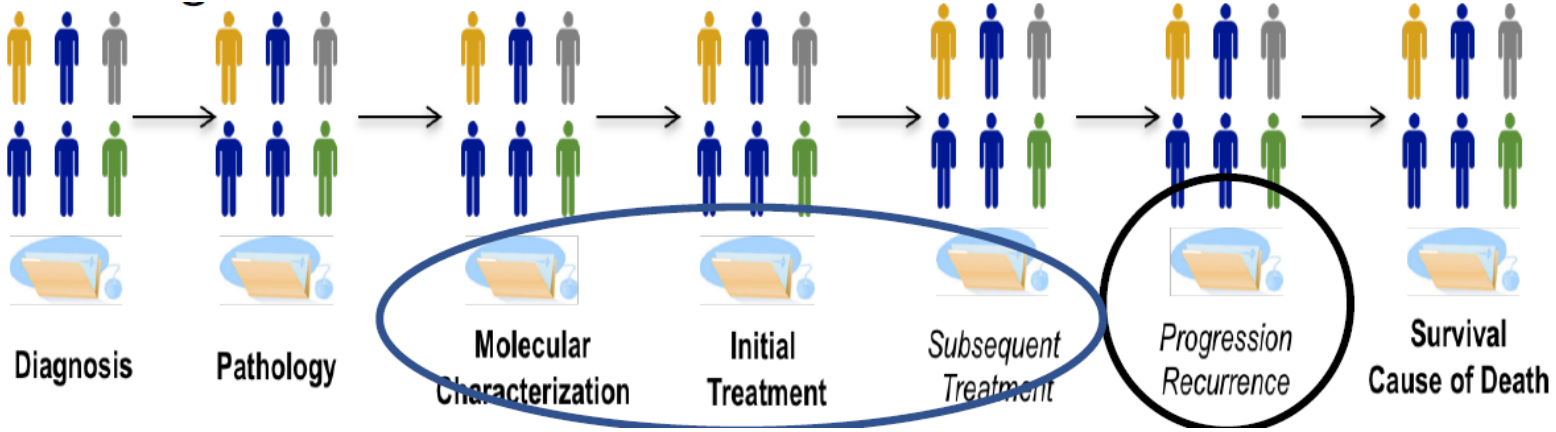
Precision Oncology Surveillance

Described last year:
Aim 1: Information capture of unstructured clinical text using Natural Language Processing (NLP) and Deep Learning algorithms
Aim 2: Information integration and analysis to understand drivers in patterns of cancer outcomes and predict clinical endpoints
Aim 3: Data-driven modeling of patient-specific and population level health trajectories

Aim 1: Advanced machine learning for scalable information extraction from unstructured clinical reports **and medical images.**
Aim 2: Scalable and visual analytics to understand the associations of patient trajectories **and the exposome** with patient outcomes and enable better **clinical trials matching**.
Aim 3: Precision data-driven modeling of patient trajectories **with a focus on cancer recurrence**

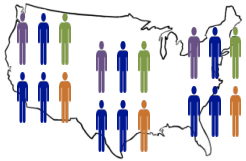


Pilot 3



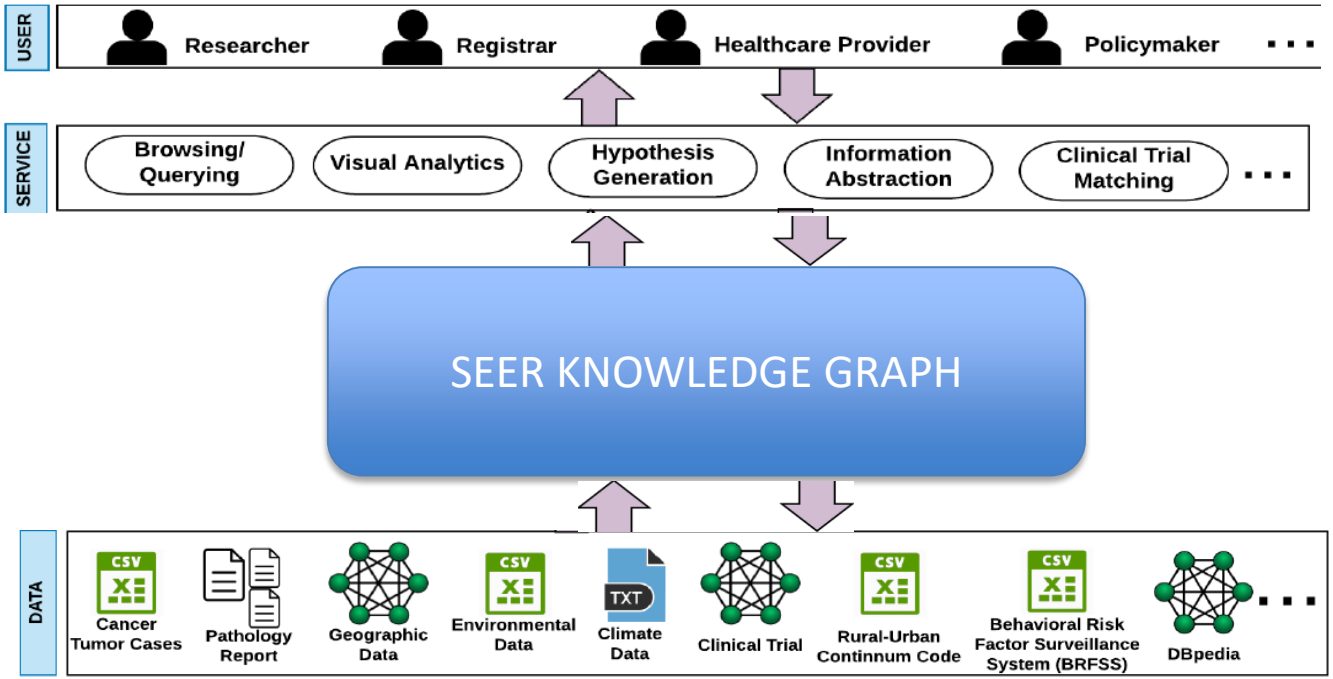
Present efforts

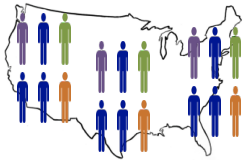
Future expansion



Pilot 3

Ultimate goal is VERY ambitious



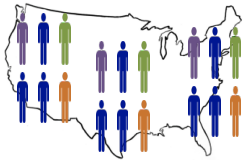


Pilot 3

The pilot has entered a virtuous cycle in which the pilot NLP developments are implemented in four registries with the goal of scaling to 35% of the US population in the next two years → capture data more efficiently and improve accuracy and timeliness by understanding when failures occur

It will include pathology images next. It may be useful to have competing groups to classify synthetic images from the SEER data base.

In years 4 and 5 it will extend the data with recurrence and biomarkers. Ultimate goal may take time to achieve, but products developed in the meantime are already improving the utility of the SEER registries.



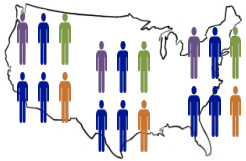
Pilot 3

Implementation of knowledge graph, building on other attributes of data such as behavioral, environmental, co-morbidities provides path towards use in clinical trial matching and ultimately individualized risk score for cancer

Some additional directions suggested by the working group:

- Distribute tools and establish collaborations with cancer centers. Convene collaborative community similar to the RAS initiative.

- Within Aim 1, apply existing and novel deep learning methods through CANDLE to pathology and medical images at scale. It will allow creation of knowledgebase with a semantic graph representation of each patient's care trajectory, life trajectory, therapeutic trajectory, and disease trajectory



Pilot 3

Within Aim 2, provide a framework and infrastructure for matching cancer patients to clinical trials. For this to be feasible, the trial has to be itself computable, and the patient phenotype does too.

Within Aim 3, build a predictive model for recurrence for all cancer patients. This is a stretch goal even for a few cancer disease areas.

Uncertainty Quantification

Cuts across all pilots. The central questions that uncertainty quantification of deep learning implementations needs to answer:

- Uncertainty of individual predictions
- Quality assessment of input data
- Model quality improvements
- Determination of quantity and quality of needed data

Uncertainty Quantification

- Models make many predictions, only some of which are certain enough to be actionable; UQ tells us which ones
- Input data has varying degrees of errors; UQ allows determining the clean subset
- UQ is essential for determining data modeling uncertainty in situations with limited data
- UQ allows one to choose models that maximize certainty
- Experiments/data are expensive; UQ allows to choose which data is most valuable
- Deep learning is important for many DOE and NCI projects, the method developed will have wide applicability.

Uncertainty Quantification

The technology driven by the pilots is innovative and interesting (e.g., abstention classifiers and the mixture models for heteroscedastic uncertainties). It also appears that the UQ effort has had impact by helping cancer researchers think about uncertainties.

The UQ plans for the 4th and 5th years are directed at the important issues. Several objectives for the coming two years appear especially challenging:

- **UQ for transfer learning:** This is essentially the application of UQ to extrapolation. Applications to ML models with their complex empirical model structure seems particularly challenging.
- **Small data:** Dealing with UQ in the small-data limit is greatly enhanced by the availability of uncertainty estimates for the data, but this is apparently not always available. in these applications. There may be fundamental limits to what can be done without it.
- **Efficient abstention:** The observed inefficiency in abstention may not be the fault of the formulation/algorithms, but rather the limitations of the available data. Again, there may be fundamental limits to what can be done.

Uncertainty Quantification

Two suggestions of the WG that may be helpful to the UQ team:

- In the application to multi-scale simulations in pilot 2, if they have not already, the team may want to consider some of the ideas underlying Bayesian optimal experimental design, as it seems that the challenges of healing coarse simulations, and balancing exploration and exploitation may have similarities with the experimental design problem.

- It seems that the proposed efforts could benefit from developments in the CS community around reliability of machine learning models, and related issues. Tracking these developments would be helpful.

Conclusions

- Overall, the three pilots have developed impressively over the past three years.
- Each pilot is helping define the quality and type of data that are necessary to make progress
- Important work has been done across the pilots on uncertainty quantification. This is essential for the utility of knowledge extracted from data and for the predictive models in all pilots.
- We have learned that the problems are, if anything, more difficult and challenging than initially imagined.
- There is a sharpening and focusing of the aims for years 4 and 5 following the lessons learned from the existing pilots.
- While I have described the progress mostly through an NCI perspective, the tools and methodologies developed in the pilots have broad applicability to many machine learning problems. Beyond the developed methodologies, the scale of the problems tackled by the pilots is helping DOE define what architectures are needed at the exascale.

The future: my take

- It is important for DOE, NCI and their respective communities to understand how the program evolves beyond the next two years.
- Before the end of Year 4 and 5, NCI and DOE should have a deep technical review of each pilot, in greater depth than is possible with the Working Group and try to determine what end-point is possible in each of the pilots. They are, after all, pilots and pilots should end and should transition to programs.
- Armed with the knowledge acquired in the pilots, NCI and the community should plan on how broad and deep the application of machine learning and deep neural networks to the many cancer problems. The pilots are already quite general, but the tools developed in the pilots are potentially of much broader utility.
- Now that we have these pilot projects for 3 years, it is important for the agencies to take advantage of this opportunity and seriously explore potential benefits of a sustained partnership – delivering computing advances for DOE and developing cutting-edge techniques for cancer research.