New York-Presbyterian and Columbia University Irving Medical Center Requirements Analysis Report

December 3rd, 2024

Disclaimer

This document was prepared as an account of work sponsored by the United States Government. While this document is believed to contain correct information, neither the United States Government nor any agency thereof, nor the Regents of the University of California, nor the Regents of the University of Texas System, nor any of their employees, makes any warranty, express or implied, or assumes any legal responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by its trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or the Regents of the University of California or the Regents of the University of Texas System. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof or the Regents of the University of California, or the Regents of the University of Texas System.

New York-Presbyterian and Columbia University Irving Medical Center Requirements Analysis Report

December 3rd, 2024

The Engagement and Performance Operations Center (EPOC) is supported by the National Science Foundation under Grant No. 2328479.

ESnet is funded by the U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research. Carol Hawk is the ESnet Program Manager.

ESnet is operated by Lawrence Berkeley National Laboratory, which is operated by the University of California for the U.S. Department of Energy under contract DE-AC02-05CH11231.

This is a University of California, Publication Management System report number LBNL-2001629[1].

[1] https://escholarship.org/uc/item/3gr7j44m

Participants & Contributors

Dr. R Graham Barr, Columbia University Irving Medical Center
James Bossio, Columbia University Irving Medical Center
Charles Corona, Columbia University Irving Medical Center
Christopher Damoci, Herbert Irving Comprehensive Cancer Center
Eme Ejike, Columbia University Irving Medical Center
Steven Erde, Columbia University Irving Medical Center
Michael Faucher, Columbia University Irving Medical Center
Pradeep R Godhala, NewYork-Presbyterian Hospital
Krystal Vega Guilbe, Columbia University Irving Medical Center
Nishith Kamdar, NewYork-Presbyterian Hospital
Dr. Despina Kontos, Columbia University Irving Medical Center
Jim Kyriannis, NYSERNet
Dr. Michal Levo, Columbia University Irving Medical Center
Wil McKoy, Columbia University Irving Medical Center
Ken Miller, ESnet
Jennifer Oxenford, NYSERNet
Zhani Pellumbi, Columbia University Irving Medical Center
B.J. Pinsky, NewYork-Presbyterian Hospital
Dr. Raul Rabadan, Columbia University Irving Medical Center
Mark Turczan, NewYork-Presbyterian Hospital
Chris Sander, NewYork-Presbyterian Hospital
William Wozniak, NewYork-Presbyterian Hospital
Jason Zurawski, ESnet

Report Editors

Jason Zurawski, ESnet: zurawski@es.net

Contents

1 Executive Summary

Deep Dive Review Purpose and Process

EPOC uses the Deep Dive process to discuss and analyze current and planned science, research, or education activities and the anticipated data output of a particular use case, site, or project to help inform the strategic planning of a campus or regional networking environment. This includes understanding future needs related to network operations, network capacity upgrades, and other technological service investments. A Deep Dive comprehensively surveys major research stakeholders' plans and processes to investigate data management requirements over the next 5–10 years. Questions crafted to explore this space include the following:

- How, and where, will new data be analyzed and used?
- How will the process of doing science change over the next 5–10 years?
- How will changes to the underlying hardware and software technologies influence scientific discovery?

Deep Dives help ensure that key stakeholders have a common understanding of the issues and the actions that a campus or regional network may need to undertake to offer solutions. The EPOC team leads the effort and relies on collaboration with the hosting site or network, and other affiliated entities that participate in the process. EPOC organizes, convenes, executes, and shares the outcomes of the review with all stakeholders.

This Review

Between February and June 2024, staff members from the Engagement and Performance Operations Center (EPOC) met with researchers and staff from New York-Presbyterian (NYP), Columbia University Irving Medical Center (CUIMC), and NYSERNet for the purpose of a Deep Dive into scientific and research drivers. The goal of this activity was to help characterize the requirements for a number of campus use cases, and to enable cyberinfrastructure support staff to better understand the needs of the researchers within the community.

Material for this event included the written documentation from each of the profiled research areas, documentation about the current state of technology support, and a write-up of the discussion that took place via e-mail and video conferencing. The case studies highlighted the ongoing challenges and opportunities that NYP and CUIMC have in supporting a cross-section of established and emerging research use cases. Each case study mentioned unique challenges which were summarized into common needs.

This review includes case studies from the following campus stakeholder groups:

- Dr. Despina Kontos, Computational Biomarker Imaging Group (CBIG)
- Dr. Michal Levo, Gene Regulation and Genome Organization
- Dr. R Graham Barr, Respiratory Epidemiology

- Dr. Raul Rabadan, Rabadan Lab
- Oncology Precision Therapeutics Imaging Core (OPTIC)
- New York-Presbyterian Hospital Technology Support
- Columbia University Irving Medical Center Cloud Services

Material for this event included the written documentation from each of the profiled research areas, documentation about the current state of technology support, and a write-up of the discussion that took place via e-mail and video conferencing.

The case studies highlighted the ongoing challenges and opportunities that NYP and CUIMC have in supporting a cross-section of established and emerging research use cases. Each case study mentioned unique challenges which were summarized into common needs.

The review produced several important findings and actions from the case studies and subsequent virtual conversations:

- Knowledgeable staff to assist researchers with tasks like data mobility, use of either on-site or cloud-based computation, or overall improvement to scientific workflows, is limited. Building in-house expertise to address these aspects of technology support would benefit a large population of users.

- Research data volumes, both of individual files as well as data sets that are produced and manipulated, will continue to grow in the coming years. The common file size is now multiple GB, data sets are TB sized, and years of research data are approaching PB size. The campuses must adapt to these needs and make storage options available.

- Data sharing is currently inefficient and relies on a heterogeneous environment that fits individual researcher approaches. The general lack of high-speed networking is a factor; however, the use of different and ineffective tools makes the issue much more challenging.

- The research community desires uniformity in the availability of computation, storage, and networking resources. Providing this would require revisiting the age and capabilities of local resources, along with what is available through external parties like cloud providers.

- An ideal data architecture should involve predictable and dependable ingest procedures for data that would capture the steps of de-identification and classification of data, methods to appropriately store and backup data sets, and ways to efficiently retrieve data for use in analysis pipelines.

- Predictable access to computation and storage resources does not have to prescribe a location. A number of use cases can function on local or remote resources efficiently. The primary concern is availability and capability: once a researcher creates an analysis pipeline for their work, they want to rely on the technology to be available and efficient to the tasks they have designed.

- Use of GPUs in AI and ML will increase across research areas. These can be available locally or in clouds.

- Cloud connectivity is supported by Mega port and is seeing increased use. Internet2 does offer Megaport connectivity options, but only to direct connectors and members of Internet2. NYP and CUIMC should consider this to simplify their WAN connectivity.

- Exploring the concept of a condo computing model for the campus is recommended. This mechanism may allow the campus to scale computing needs against what funding is available from individual research groups. Resources are allocated to individuals initially based on purchase power but can also be used by others when available.

- The ability to offer a campus-wide seamless backup system would be desirable, as they have no formalized way to manage this now resulting in individuals approaching this in different ways.

- Create a campus-wide data architecture.

- Establishing application-level monitoring within and external to the networks using perfSONAR testing.

- Allowing visibility into traffic with passive taps and analysis frameworks to prevent friction of data transfer.

- Create unified storage enclaves. Investing in institutional storage and making it available to instruments that require this capability, will ensure that research data has a unified ingress and egress.

- Creating and supporting a common data mobility platform.

2 Deep Dive Findings & Actions

The deep dive process helps to identify important facts and opportunities from the profiled use cases. The following outlines a set of findings and actions from the NYP and CUIMC Deep Dive that summarize important information gathered during the discussions surrounding case studies, and possible ways that could improve the CI support posture for the campus.

2.1 Findings

- Research data volumes, both of individual files as well as data sets that are produced and manipulated, will continue to grow in the coming years. The common file size is now multiple GB, data sets are TB sized, and years of research data are approaching PB size. The campuses must adapt to these needs and make storage options available.

- Data sharing is currently inefficient and relies on a heterogeneous environment that fits individual researcher approaches. The general lack of high-speed networking is a factor; however, the use of different and ineffective tools makes the issue much more challenging. Web and cloud-based sharing methods (e.g., Dropbox, etc.) are widely used, but break down as the file sizes and data sets grow. Shipping media is still widely used. Relying on "neutral" sharing locations (e.g., AWS) work well when commutation is involved, but do not scale as just a sharing mechanism. For the later to be a solution, entire platforms must be built to offer storage, computation, and access control: this can be expensive and may be out of reach for some research efforts. Care must be taken to ensure that efforts to improve sharing are working in concert with efforts to classify, deidentify, and protect data.

- The research community desires uniformity in the availability of computation, storage, and networking resources. Providing this would require revisiting the age and capabilities of local resources, along with what is available through external parties like cloud providers. It is feasible to design a system that would facilitate a uniform (e.g., containerized) approach that would scale to local and remote use cases.

- An ideal data architecture (assuming no limitations on cost) should involve predictable and reliable ingest procedures for data that would capture the steps of de-identification and classification of data, methods to appropriately store and backup data sets (as well as applying permissions for who can access), and ways to efficiently retrieve data for use in analysis pipelines.

- Predictable access to computation and storage resources does not have to prescribe a location. Several use cases can function on local or remote resources efficiently. The primary concern is availability and capability: once a researcher creates an analysis pipeline for their work, they want to rely on the technology to be available and efficient to the tasks they have designed. In general, they do not want to build and operate their own computational system (more time working on technology takes away from the ability to do research). In particular, "interactive computing" (e.g., the use of notebook software) could be a valuable way to approach future analysis tasks.

- On-site computational resources can be a challenge to maintain due to requirements and cost. Cloud computing vendors can scale their offerings, and offer access to cutting edge technology, while the campus may need to utilize hardware over a longer cycle with minimal refresh. Having both options available for researchers is recommended.

- There is a fundamental friction with the use of off-site cloud resources for medical work. Historically the security concerns have favored "keeping the data close", which would imply the clinical environment is better served to maintain its own hardware and software and absorb the costs to support the researchers.

- Overall, cloud use can be seen to focus on these use cases:
    - Backups
    - AI/ML experimentation
    - Creating micro-services
    - Analytics
    - File Sharing

- Use of GPUs in AI and ML will increase across research areas. These can be available locally or in clouds.

2.2 Actions

- Knowledgeable staff to assist researchers with tasks like data mobility, use of either on-site or cloud-based computation, or overall improvement to scientific workflows, is limited. Building in-house expertise to address these aspects of technology support would benefit a large population of users.

- Limited bandwidth, and lack of adequate data transfer tools, is causing friction for researchers looking for data mobility. Dr. Kontos and Dr. Levo both have experienced issues with transferring data from previous institutions and can benefit from dedicated data movement hardware and software. Both can use support going forward with how they can effectively use the Globus service that has been implemented.

- Cloud connectivity is supported by Megaport and is seeing increased use. Internet2 does offer Megaport connectivity options, but only to direct connectors and members of Internet2. NYP and CUIMC should consider this to simplify their WAN connectivity.

- Exploring the concept of a condo computing model for the campus is recommended. This mechanism may allow the campus to scale computing needs against what funding is available from individual research groups. Resources are allocated to individuals initially based on purchase power but can also be used by others when available.

- The ability to offer a campus-wide seamless backup system would be desirable, as they have no formalized way to handle this now resulting in individuals approaching this in different ways.

- Use of scientific instruments (e.g., MRIs, microscopes, etc.) still relies on a way to facilitate "remote access" to some on-premises resources. For example, a number of analysis tools must be run on the machines that interact with the instruments, due to software licensing requirements. Building up the ability of remote users to safely and securing access these instruments should be a priority.

- Create a campus-wide data architecture. This should be done by first identifying where data must flow. In some cases, this may be “east and west” (e.g., within the campus), and in others it may be “north and south” (e.g., egressing or ingressing campus). From the use cases, there are a number that fit both definitions. To accomplish data mobility in either case, the technology group can leverage the network in the following ways:
    - Existing network paths and optimizing transfers.
    - Creating new and/or dedicated network paths

- Establishing application-level monitoring within and external to the networks using perfSONAR testing, located at several places in and around campus.

- Allowing visibility into traffic with passive taps and analysis frameworks to prevent friction of data transfer.

- Create unified storage enclaves. Storage on campus is a mixture of private solutions and some institutional capabilities. Investing in institutional storage and making it available to instruments that require this capability, will ensure that research data has a unified ingress and egress. A campus-supported SAN will also allow researchers the option to purchase access and be assured that sensitive data is being handled appropriately.

- Creating and supporting a common data mobility platform. Data transfer outside of the environment requires dedicated hardware and software. Creating a set of Data Transfer Nodes (DTNs) that have the capability to send 10Gbps+ streams using effective transfer tools.

- Common ways to egress the network that leverages the R&E and cloud connectivity options. If a Clinical DMZ is established, it should have a way to leverage WAN connectivity directly. This can be a short hop to the border, or a direct connection to NYSERNET, Internet2, or clouds (e.g., Megaport).

3 Process Overview and Summary

3.1 Campus-Wide Deep Dive Background

Over the last decade, the scientific community has experienced an unprecedented shift in the way research is performed and how discoveries are made. Highly sophisticated experimental instruments are creating massive datasets for diverse scientific communities and hold the potential for new insights that will have long-lasting impacts on society. However, scientists cannot make effective use of this data if they are unable to move, store, and analyze it.

The Engagement and Performance Operations Center (EPOC) uses the Deep Dives process as an essential tool as part of a holistic approach to understand end-to-end research data use. By considering the full end-to-end research data movement pipeline, EPOC is uniquely able to support collaborative science, allowing researchers to make the most effective use of shared data, computing, and storage resources to accelerate the discovery process.

EPOC supports five main activities.

- Roadside Assistance via a coordinated Operations Center to resolve network performance problems with end-to-end data transfers reactively.
- Application Deep Dives to work more closely with application communities to understand full workflows for diverse research teams in order to evaluate bottlenecks and potential capacity issues.
- Network Analysis enabled by the NetSage monitoring suite to proactively discover and resolve performance issues.
- The Data Mobility Exhibition and associated work with our simplified portal to check transfer times against known performant end points;
- Coordinated Training to ensure effective use of network tools and science support.

Whereas the Roadside Assistance portion of EPOC can be likened to calling someone for help when a car breaks down, the Deep Dive process offers an opportunity for broader understanding of the longer term needs of a researcher. The Deep Dive process aims to understand the full science pipeline for research teams and suggest alternative approaches for the scientists, institutional IT support, and national networking partners as relevant to achieve the long-term research goals via workflow analysis, storage/computational tuning, identification of network bottlenecks, etc.

The Deep Dive process is based on an almost 20-year practice used by ESnet to understand the growth requirements of Department of Energy (DOE) facilities[2]. The EPOC team adapted this approach to work with individual science groups through a set of structured data-centric conversations and questionnaires.

[2] https://fasterdata.es.net/science-dmz/science-and-network-requirements-review

3.2 Campus-Wide Deep Dive Structure

The Deep Dive process involves structured conversations between a research group and relevant IT professionals to understand at a broad level the goals of the research team and how their infrastructure needs are changing over time.

The researcher team representatives are asked to communicate and document their requirements in a case-study format that includes a data-centric narrative describing the science, instruments, and facilities currently used or anticipated for future programs; the advanced technology services needed; and how they can be used. Participants considered three timescales on the topics enumerated below: the near-term (immediately and up to two years in the future); the medium-term (two to five years in the future); and the long-term (greater than five years in the future).

The case study process tries to answer essential questions about the following aspects of a workflow:

- ***Research & Scientific Background***—an overview description of the site, facility, or collaboration described in the Case Study.
- ***Collaborators***—a list or description of key collaborators for the science or facility described in the Case Study (the list need not be exhaustive).
- ***Instruments and Facilities: Local & Non-Local***—a description of the network, compute, instruments, and storage resources used for the science collaboration/program/project, or a description of the resources made available to the facility users, or resources that users deploy at the facility or use at partner facilities.
- ***Process of Science***—a description of the way the instruments and facilities are used for knowledge discovery. Examples might include workflows, data analysis, data reduction, integration of experimental data with simulation data, etc.
- ***Computation & Storage Infrastructure: Local & Non-Local***—The infrastructure that is used to support analysis of research workflow needs: this may be local storage and computation, it may be private, it may be shared, or it may be public (commercial or non—commercial).
- ***Software Infrastructure***—a discussion focused on the software used in daily activities of the scientific process including tools that are used locally or remotely to manage data resources, facilitate the transfer of data sets from or to remote collaborators, or process the raw results into final and intermediate formats.
- ***Network and Data Architecture***—description of the network and/or data architecture for the science or facility. This is meant to understand how data moves in and out of the facility or laboratory focusing on local infrastructure configuration, bandwidth speed(s), hardware, etc.
- ***Resource Constraints***—non-exhaustive list of factors (external or internal) that will constrain scientific progress. This can be related to funding, personnel, technology, or process.

- ***Outstanding Issues*—**Listing of any additional problems, questions, concerns, or comments not addressed in the aforementioned sections.

At a physical or virtual meeting, this documentation is walked through with the research team (and usually cyberinfrastructure or IT representatives for the organization or region), and an additional discussion takes place that may range beyond the scope of the original document. At the end of the interaction with the research team, the goal is to ensure that EPOC and the associated CI/IT staff have a solid understanding of the research, data movement, who's using what pieces, dependencies, and time frames involved in the Case Study, as well as additional related cyberinfrastructure needs and concerns at the organization. This enables the teams to identify bottlenecks or areas that may not scale in the coming years, and to pair research teams with existing resources that can be leveraged to more effectively reach their goals.

3.3 NYP and CUIMC Deep Dive Background

Between February and June 2024, EPOC organized a Deep Dive in collaboration with NYP and CUIMC to characterize the requirements for several key science drivers. The representatives from each use case were asked to communicate and document their requirements in a case-study format.   These included:

- Dr. Despina Kontos, Computational Biomarker Imaging Group (CBIG)
- Dr. Michal Levo, Gene Regulation and Genome Organization
- Dr. R Graham Barr, Respiratory Epidemiology
- Dr. Raul Rabadan, Rabadan Lab
- Oncology Precision Therapeutics Imaging Core (OPTIC)
- NewYork-Presbyterian Hospital Technology Support
- Columbia University Irving Medical Center Cloud Services

3.4 Organizations Involved

The Engagement and Performance Operations Center (EPOC) was established in 2018 as a collaborative focal point for operational expertise and analysis and is jointly led by the Texas Advanced Computing Center (TACC) and the Energy Sciences Network (ESnet). EPOC provides researchers with a holistic set of tools and services needed to debug performance issues and enable reliable and robust data transfers. By considering the full end-to-end data movement pipeline, EPOC is uniquely able to support collaborative science, allowing researchers to make the most effective use of shared data, computing, and storage resources to accelerate the discovery process.

The Energy Sciences Network (ESnet) is the primary provider of network connectivity for the U.S. Department of Energy (DOE) Office of Science (SC), the single largest supporter of basic research in the physical sciences in the United States. In support of the Office of Science programs, ESnet regularly updates and refreshes its understanding of the networking requirements of the instruments, facilities, scientists, and science programs that it serves. This focus has helped ESnet to be a highly successful enabler of scientific discovery for over 25 years.

The Texas Advanced Computing Center (TACC) at the University of Texas at Austin designs and deploys the world's most powerful advanced computing technologies and innovative software solutions to enable researchers to answer complex questions to help them gain insights and make discoveries that change the world. TACC's environment includes a comprehensive cyberinfrastructure ecosystem of leading-edge resources in high performance computing (HPC), visualization, data analysis, storage, archive, cloud, data-driven computing, connectivity, tools, APIs, algorithms, consulting, and software.

NewYork-Presbyterian is one of the nation’s most comprehensive, integrated academic healthcare systems, dedicated to providing the highest quality, most compassionate care and service to patients in the New York metropolitan area, nationally, and around the world. In collaboration with two renowned medical schools, Weill Cornell Medicine and Columbia University Vagelos College of Physicians and Surgeons, NewYork-Presbyterian is consistently recognized as a leader in innovative, patient-centered clinical care, research and medical education.

Columbia University Irving Medical Center Columbia University Irving Medical Center (CUIMC) is a clinical, research, and educational enterprise located on a campus in northern Manhattan. We are home to four professional colleges and schools that provide global leadership in scientific research, health and medical education, and patient care.

NYSERNet is a non-profit organization. Our mission is to advance the research and educational missions of our members by delivering a full range of customized, progressive, and affordable end-to-end data and networking technology solutions.

The networks we build with technology are only as effective as our community network. This is why we consistently cultivate opportunities to collaborate through our widely accessible professional development conferences and events.

4 NYP and CUIMC Case Studies

NYP and CUIMC presented a number use cases during this review. These are as follows:

- Dr. Despina Kontos, Computational Biomarker Imaging Group (CBIG)
- Dr. Michal Levo, Gene Regulation and Genome Organization
- Dr. R Graham Barr, Respiratory Epidemiology
- Dr. Raul Rabadan, Rabadan Lab
- Oncology Precision Therapeutics Imaging Core (OPTIC)
- NewYork-Presbyterian Hospital Technology Support
- Columbia University Irving Medical Center Cloud Services

Each of these Case Studies provides a glance at research activities, the use of experimental methods and devices, the reliance on technology, and the scope of collaborations. It is important to note that these views are primarily limited to current needs, with only occasional views into the event horizon for specific projects and needs into the future. Estimates on data volumes, technology needs, and external drivers are discussed where relevant.

4.1 Dr. Despina Kontos, Computational Biomarker Imaging Group (CBIG)

Content in this section authored by Dr. Despina Kontos, Columbia University Irving Medical Center (CUIMC)

4.1.1 Use Case Summary

Dr. Kontos is a computer scientist with training in radiologic imaging, clinical epidemiology, biostatistics, and cancer biology. Prior research includes medical image analysis, machine learning, and artificial intelligence with a focus on cancer imaging. The Kontos Lab is located within the Center for Innovation in Imaging Biomarkers and Integrated Diagnostics (CIMBID) in the Vagelos College of Physicians and Surgeons (VP&S) at Columbia University

The lab's research vision is to function as a translational catalyst between imaging data science and clinical cancer research by integrating image analysis, machine learning, and artificial intelligence in clinically relevant cancer imaging applications. The lab has developed a rigorous research program investigating the role of novel quantitative imaging biomarkers for improving personalized decisions for cancer screening, prognosis, and treatment. Research interests also include leveraging data science and artificial intelligence to integrate imaging biomarkers with non-imaging data, such as electronic health record (EHR) data, genetics, histopathology, and genomic data, into integrated precision diagnostics.

4.1.2 Collaboration Space

The lab collaborates with faculty in the Herbert Irving Cancer Center (HICC), the Department of Biomedical Informatics (DBMI), the Mailman School of Public Health, the Fu Foundation School of Engineering and Applied Science, and the Data Science Institute. The lab has also established a network of multi-disciplinary national and international collaborations, including a range of experts in biomedical informatics, bioinformatics, medical physics, radiology, radiation oncology, genetics, cancer biology, pathology, oncology, biostatistics, and epidemiology, and have led/co-lead large research consortia on evaluating novel multi-modal cancer imaging biomarkers in diverse populations.

4.1.3 Instruments & Facilities

4.1.3.1 Department of Radiology Information Technology and Image Analysis Resources

The Department of Radiology at Columbia University has a 165TB Philips Vue PACS system, 80-90 reading workstations and EPIC EHR system. The Nuance PowerScribe 360 dictation system is used by radiologists for generating diagnostic reports for distribution throughout the enterprise and association. All reports are generated using real-time speech-to-text translation, which permits signing of reports upon completing the reading of an exam. The dictation client is integrated with PACS by means of sharing login, patient, and exam information.

The Department of Radiology also has a cutting-edge post-processing 3D center for advanced image processing and analysis (CAIPA) lab providing the latest advanced quantitative and qualitative imaging. CAIPA is equipped with both server and workstation solutions for image processing and analysis, including 5 Vital workstations, GE AW server, Tera recon server, Vital Image server and Invivo Dynsuite Neuro and Invivo DynaCAD server.

4.1.3.2 AWS infrastructure

The AWS infrastructure consists of EC2 instances for compute, S3 object storage for persistent storage, AWS Parallel Cluster for Slurm HPC workloads, and EFS or similar service for a shared cluster filesystem for use during active processing. S3 object storage is used for long-term storage of data. S3 tiering is used to reduce costs of infrequently used data or data that needs to be archived. S3 is made accessible to both institutions with appropriate access privileges and logging in place. A shared filesystem, such as EFS, is used for the Slurm cluster. Data is copied from S3 into the shared filesystem, processed, analyzed, and copied back to S3 for final storage. The shared filesystem isn't used for long-term data storage even though the filesystem is not ephemeral like the cluster worker nodes. Batch workloads are run on AWS Parallel Cluster which provides a familiar Slurm environment for workload orchestration. The cluster head node is the only persistent compute resource and is sized as small as possible, as its sole purpose is to orchestrate the jobs. There are at least two different queue types, one containing regular CPU compute instances, and one with GPU capabilities. Other queues may be configured if compute, memory, and GPU requirements change. Worker nodes in the queues will be spun up on-demand as jobs are submitted to the cluster. When no jobs are running the worker nodes will automatically terminate. Persistent data will be stored in S3 or on a shared filesystem. Interactive workloads are run on regular EC2 instances, which can be turned on and off on-demand. Services such as Amazon SageMaker may be used in lieu of EC2 instances for some AI/ML work to reduce operational overhead of maintaining custom systems.

4.1.3.3 Deep learning Medical Physics Lab Computational Resources

As part of the Department of Radiology, the team has access to several high-end deep learning systems. Through the Department of Radiology, the lab has access to state-of-the-art rackmount 4U Rackmount HGX A100 GPU System with 8x80 GB A100 GPUs with 1TB of system memory and 25TB of NVMe storage for running large deep learning algorithms. In addition, the medical physics lab also has high end GPU processing cards- 13 GPU cards (4x 12GB Nvidia GeForce Titan Xp, 5x 1080 Nvidia GeForce Ti GPU's, 4x 24GB Titan RTX with NVlink). Additionally, the team currently also has a NVIDIA DGX station for leading-edge AI development projects (4x32GB NV linked Tesla V100 for running large (up to 128GB) GPU memory problem).

For parallel image analysis computation, a system with multi-CPU 36 core Xeon processor with 256 GB RAM system is available in the lab as well. All systems run

Ubuntu Linux with shared 200 TB of RAID10 data drives and have Nvidia CUDA 11 library with latest deep learning analysis software such as tensorflow, pytorch and keras installed. In addition, all team members have access to clinical workstations, image analysis servers and research storage space that interface with all hospital research systems, including PACS.

4.1.3.4 Deep Learning Computational Biology and Bioinformatics Resources

C2B2 is a Computing CORE managed by DSBIT, the largest research computing facility on the CUIMC campus, located in the ICRC building. Its Data Center has 100 high-density racks, powered by one MW of battery-backed UPS. It offers CPU+GPU based High Performance Computing Cluster, HIPAA compliant enterprise grade data storage, high bandwidth interconnects. Policy of no-profit and subsidies make it most affordable facility in New York. The C2B2 HPC system is on the 2013 Top500 list of supercomputers worldwide.

The team will have access to the High-Performance Computing Environment (HPCE) of the Center for Computational Biology and Bioinformatics (C2B2) which offers access to several High-Performance Computing (HPC) systems, including multiple high performance compute clusters as well as high-memory systems.
The HPC cluster consists of 12,000 CPU cores in the latest AMD EPYC processors, over 1M CUDA cores in NVIDIA 40x L40S/48GB GPUs, one NVIDIA Superchip GH200 with ARM/480GB+H100/96GB GPU, 64 compute nodes -- each with large memory (i.e., 768GB to 1.5TB), as well as pre-installed software (GCC, Python, R, MPI, MATLAB, BLAST). The entire HPC cluster is interconnected with a 10GB mesh network. And for highly coupled parallel computations, a portion of the cluster is equipped with a HDR200/100 infiniband network. Cluster job scheduling is managed with Univa Grid Engine. Among the wide range of scientific and computational software available are, the latest GNU and Intel compilers for C and Fortran, Perl interpreters, Java SDKs, Matlab, BLAST, EMBOSS, HMMER, MUMmer, clustalW, PAML, PHYLIP, BioConductor, Phred and Phrap, GeneHunter, Fastlink, Merlin, PDT, TRANSMIT, Pseudomarker, Analyze, Autosacan, GOLD, plus many other utilities and programs.

4.1.3.5 Additional Resources

Dual HP blade servers house the virtualization infrastructure used for development and production systems. The virtualization infrastructure hosts web, application, database, and infrastructure management systems. Microsoft's Active Directory (AD) provides single sign-on secure authentication to the entire computing environment.

4.1.4 Data Narrative

All investigators recognize the importance of data sharing and management. To execute this data sharing and management plan, the MPIs will utilize two resources:

1) Columbia University Library Academic Commons
2) A dedicated GitHub repository for this project.

Columbia University Academic Commons is a publicly accessible digital repository that provides open, persistent access to research produced at Columbia University. The Columbia University Libraries manage Academic Commons and are part of the Libraries' long-term digital storage system, which ensures that files are replicated and stored in at least two distinct locations. The repository follows findable, accessible, interoperable, and reusable (FAIR) data principles and uses unique, persistent identifiers and rich metadata to enhance the accessibility of each work.

4.1.4.1 Data Sharing and Management Terms

4.1.4.1.1 Element 1: Data Type

A. Types and amount of scientific data expected to be generated in the project:
In this proposed project, clinical retrospective databases of matching digital breast tomosynthesis data and patient health data will be created at Columbia University Irving Medical Center (CUIMC) and the University of Pennsylvania (UPenn). These databases will be stored and managed by each site's clinical data teams under the protection of each site's Human Subjects Protection Protocols. Following NIH DMS policy, as the primary imaging and clinical data are not generated prospectively specifically for the purposes of this study, the raw clinical data will not be shared. But all research-generated imaging measurements, codes, and trained AI models will be publicly accessible. We will also make every effort to work with both CUIMC and UPenn to make deidentified original imaging data available under appropriate institutional permissions via NCI's TCIA.

The project will create multi-center data of tomosynthesis imaging and clinical health medical variables from approximately 250,000 women. The total size of the data collected is projected to be 750 terabytes.

We expect to generate the following data file types and formats during this project:

1. Derived imaging features (tomosynthesis images): DICOM, xml, jpg, png, json
2. Derived imaging feature data: cvs, xlsx
3. AI-modeling software code: Python libraries, Jupyter Notebooks, Keras, Tensorflow, GitHub

There is no genomic data involved.

B. Scientific data that will be preserved and shared, and the rationale for doing so:
Derived imaging features, AI models, and software code, and related documentations will be shared via a dedicated GitHub repository and Academic Commons. The Principal Investigators (MPIs) will share data of sufficient quality to facilitate the validation and replication of research findings outlined in the project aims. This will encourage other research groups to build upon our results and utilize the derived data for additional analyses we may not have considered. While the primary retrospective datasets will not be shared, all new derived features and

models generated from these images will be publicly accessible through GitHub and Academic Commons. All data will be de-identified prior to uploading, in compliance with privacy regulations.

C. Metadata, other relevant data, and associated documentation:
A README file and data dictionary will be generated and deposited into all the online repositories along with all shared datasets to facilitate the interpretation and reuse of the data. The README file will include method description, software pipeline settings, AI model codes, and all software algorithms for image processing and analysis. The data dictionary will define and describe all variables in the dataset.

4.1.4.1.2 Element 2: Related Tools, Software and/or Code:

The software products developed in this project, including explainable and generalizable deep learning/AI algorithms trained on our data commons, will be made publicly available for use by the research and public health communities (via our servers and GitHub account(s)). This software will be developed and shared using the latest PyTorch distributions, open-source platforms designed as a front-end for deep learning experiments that will enable easy prototyping and deployment into different operating systems. The imaging data will be analyzed with custom Python code using several libraries including PyTorch for deep learning libraries (see also above). GitHub will be the main place for sharing the code/software/tools internally and externally.

4.1.4.1.3 Element 3: Standards:

Imaging data will be in DICOM (de-identified) format mainly. Segmentation data will be in DICOM, Nifti, and xml formats. Numerical data including labels will be stored using Microsoft excel or RedCAP (password protected).

4.1.4.1.4 Element 4: Data Preservation, Access, and Associated Timelines

A. Repository where scientific data and metadata will be archived:
Our data (de-identified) will be deposited in the research database of Columbia University Academic Commons and Github and will be available to peers after the manuscripts are published. All the data will be shared with both repositories starting 12 months after the award begins and will be deposited every six months thereafter following the submission dates.

B. How scientific data will be findable and identifiable:
The data will be findable for the research community via the Columbia University Academic Commons (via their unique identifiers) when this application is funded.

We will also create data DOI (digital object identifier) as a part of our publications, allowing research community easy access to the exact data used in the publications.

C. When and how long the scientific data will be made available:
Data will be shared to the research community after the main manuscripts are published. No end date is considered for the scientific data.

4.1.4.1.5 Element 5: Access, Distribution, or Reuse Considerations

A. Factors affecting subsequent access, distribution, or reuse of scientific data: Following NIH DMS policy, as the primary imaging and clinical data are not generated prospectively for the study, the original raw clinical data will not be shared. That said, we will make every effort, pending institutional approvals, to share de-identified imaging data, or sub-sets of the dataset, in NCI's TCIA.

B. Whether access to scientific data will be controlled:
To request access of the data, researchers will be asked to sign a "data sharing agreement". Then, the data covered by the "agreement" can be accessed from our secure servers.

C. Protections for privacy, rights, and confidentiality of human research participants: Only de-identified data will be shared. All participants of the present study will be assigned a unique study identifier. Individual names and other private information will ultimately be removed from the study database, and only the unique study identifier will be used to distinguish the participants in the database. The collected data will be maintained in locked computer files and file cabinets to which only the authorized study investigators have access.

4.1.4.1.6 Element 6: Oversight of Data Management and Sharing:

Both CUIMC and UPenn have clinical data management teams that have a sharing plan compliance system as part of their quality and safety oversight. Additionally, each site's IRB will also oversee data sharing and transferring.

4.1.4.1 Data Volume & Frequency Analysis

Currently the Kontos Lab produces TB scales of on a daily basis.

4.1.4.2 Data Sensitivity

Some of the data that the Kontos Lab produces can be considered sensitive, but there are appropriate controls in place for dealing with this data.

4.1.4.3 Future Data Volume & Frequency Analysis

The Kontos Lab anticipates producing TB scales of data on a daily basis in the future.

4.1.5 Technology Support

4.1.5.1 Software Infrastructure

DSBIT supports a wide range of scientific and computational software including the latest GNU and Intel compilers for C, Fortran, python, perl interpreters, Java SDKs, Matlab, R, and BLAST.

4.1.5.2 Network Infrastructure

An Arista core switch provides a 10 GB/s backbone to the DSB data center networks. The core switch is connected by two 10 GB/s uplinks to the internet and one 10 GB/s uplink to Internet2 – for a total of 30 GB/s bandwidth to the internet.

Data center servers are assigned 1 GB/s links and the HPC cluster is equipped with a direct 10 GB/s connection to the core switch. Perimeter firewalls protect the data center networks, while load balancers distribute network traffic to the web and compute infrastructure.

4.1.5.3 Computation and Storage Infrastructure

The Department of Radiology at Columbia University maintains a 32 node Rocks cluster with 8 cores per node and 12 GB of RAM/node for in-house image processing, statistical analysis, and image registration software development. The cluster has image and statistical analysis software's installed such as: Matlab, R, ImageJ, neuroimaging packages (FSL,Freesurfer, AFNI, SPM) and image registration software like ANTS (advanced normalization tools) and ITK (image registration toolkit) .

The CBIG Lab has additional capabilities:

- Each CBIG member has a personal research laptop (Intel Core i7-1365 @ 1.80 GHz, 32 GB Memory, 1 TB storage capacity). Each workstation is equipped with wide screen monitor (add specs).
- Two GPU workstations located within the lab, each with 2x 16 core CPUs accompanied by two 64 GB NVIDIA T400 GPUs, which can be used for initial code verification and proof of concept testing.
- Two medium tier GPU servers providing faster per clock CPU cores, 128 GB of DDR5 ECC memory and dial NVIDIA RTX A4000 GPUs.

All workstations have additional 1 TB of external storage disk space provided via Microsoft One Drive. The space is password protected and requires two-factor authentication to access and is routinely backed up. Available software includes advanced image analysis and software development packages, such as Python, R, Matlab (Mathworks Inc.), SPSS (IBM), VisualStudio and .NET Framework (Microsoft), ITK toolbox (Insight), ITK-SNAP, CapTK, MIPAV, MEVIS Lab SDK, SQL Server (Microsoft), and office productivity tools.

CBIG has access with established membership to the High-Performance Computing Environment (HPCE) of the Center for Computational Biology and Bioinformatics (C2B2) which offers access to several High-Performance Computing (HPC) systems, including multiple high performance compute clusters as well as high-memory systems. The HPC cluster consists of 6,384 CPU cores, 39TB of RAM and 148 NVIDIA GPUs providing an additional 75,776 CUDA cores. The entire HPC cluster is interconnected with a 10GB mesh network. And for highly coupled parallel computations, a portion of the cluster is equipped with a 40GB QDR infiniband network. Cluster job scheduling is managed with Univa Grid Engine. A 5.5 PB enterprise-grade storage system provides high-speed, redundant storage that is tightly integrated with the HPC cluster to support big data analyses. A disk-to-disk replication storage system is used for long- term backup. And a Scalar I2k 300 slot tape robot is used for off-site disaster recovery (DR) backup. CBIG has additional entry level and medium tier options for compute and GPU processing locally within the lab. At the entry level 2x 16 Cores CPUs accompanied by 2 64GB NVIDIA T400 GPUs can be used for testing and initial code verification and proof of concept. At the medium tier options for the systems jump up significantly providing faster per clock CPU cores, 128GB of DDR5 ECC memory & dual (2) NVIDIA RTX A4000's. CBIG also has access to Amazon Web Services (AWS) Deep Learning Amazon Machine Images (DLAMI). DLAMI offers a variety of instance types ranging from small single-CPU instances to large multi-GPU instances, all of which come preconfigured with NVIDIA CUDA, NVIDIA cuDNN, and the latest releases of most popular deep learning frameworks. DLAMI images can be spun up on demand with up to 8 NDIVIA Tesla V100 GPUs, 8 NVIDIA Tesla A100 GPUs, 4 NVIDIA Tesla M60 GPUs, 4 NVIDIA T4 GPUs, or 8 NVIDIA A10G GPUs. Arm-based AWS Graviton2 processors are also available.

A 5.5 PB enterprise-grade storage system provides high-speed, redundant storage that is tightly integrated with the HPC cluster to support big data analyses. A disk-to-disk replication storage system is used for long- term backup. And a Scalar I2k 300 slot tape robot is used for off-site disaster recovery backup.

4.1.5.4 Data Transfer Capabilities

The majority of data transfer occurs within the confines of the NYP and CUIMC networks, and uses traditional tools (cloud-based, or computer-to-computer synchronization).

4.1.6 Internal & External Funding Sources

Dates Active	***Title***	***Agency***	***PI***	***Funding***	***Effort***	***Role***
7/2023-6/2028	Privacy-Aware Federated Learning for Breast Cancer Risk Assessment	NCI/NIH/DHHS	S. Bakas	$573,474/annual direct costs	10%	Co-Investigator
4/2023-3/2028	Evaluation of Novel Tomosynthesis Density Measures in Breast Cancer Risk Prediction	NIH	V. Celine, D. Kontos	$4,040,572/annual direct costs	10%	MPI
1/2023-12/2027	Genetic and radiomic markers to guide supplemental screening for breast cancer	American Cancer Society	A. McCarthy	$200,000/annual direct costs	10%	Co-Investigator
8/2022-6/2026	Next-Generation Tomosynthesis Pilot Study	NIH	R. J. Acciavatti	$224,175 /annual direct costs	4%	Collaborator

8/2022-4/2026	Framework for radiomics standardization with application in pulmonary CT scans	NIH	J. G. Gang	$288,269/annual direct costs	5%	Collaborator
8/2022-7/2025	Imaging and Liquid Biopsy Biomarkers for Precision Screening and Early Detection of Lung Cancer	V Foundation for Cancer Research	D. Kontos	$183,000/annual direct costs	5%	PI
7/2022-6/2027	MRI Radiomic Signatures of DCIS to Optimize Treatment	NHI/NCI	H. Rahbar, D. Kontos	$246,017/annual direct costs	11%	MPI
7/2023-6/2027	Combining volumetric breast density and polygenic risk scores to improve breast cancer risk assessment for black and white women	American Cancer Society	A. McCarthy	$150,000/annual direct costs	5%	Co-Investigator
6/2021-5/2026	Predictive and Diagnostic Radiomic Signatures in Non-Small Cell Lung Cancer (NSCLC) on Immunotherapy	NHI/NCI	D. Kontos, S. Katz	$451,247/annual direct costs	15%	PI
7/2020-7/2025	Medical Imaging and Data Consortium (MIDC): Rapid Response to COVID-19 Pandemic	NIH/NIBIB	M. Giger	$50,000/annual direct costs	3%	Site-PI
9/2019-6/2025	Data and Information Integration for Risk Prediction in the Era of Big Data	NIH	J. Chen	$384,725/annual direct costs	10%	Co-Investigator
9/2019-8/2024	Understanding the biological basis for the association between parenchymal texture features and breast cancer risk	NIH	S. Nyante	$87,043/annual direct costs	3%	Site-PI
7/2018-5/2024	Radiogenomic Biomarkers of	NIH	D. Kontos	$414,065/annual direct costs	8%	PI

	Breast Cancer Recurrence					

4.1.7 Resource Constraints

All these aspects are major constraints in future research, including insufficient data transfer performance, insufficient storage system space or performance characteristics, difficulty finding or accessing data in community data repositories, or unmet computing needs. However, one of the biggest constraints is inefficiency in navigating the NYP/CUIMC aspects of data governance, IT infrastructure, network bandwidth, and clearances.

4.1.8 Ideal Data Architecture

This blueprint outlines the establishment of an advanced imaging data infrastructure, tailored for NYPH, Columbia, and Cornell. The infrastructure is designed to support the management and analysis of approximately 1600,000 annual imaging studies, facilitating cutting-edge AI research. Flywheel Discovery will be used as the common backend for imaging metadata indexing and curation across all NYPH hospitals, while Flywheel Core will be implemented independently at Columbia and Cornell for research and development, with individual IRBs for collaboration on combined data. The existing high-end HPC CPU clusters at Columbia and Cornell can be leveraged to enhance the computational capacity of the infrastructure.

4.1.8.1 Hardware, Software, and Implementation Strategy

- Hardware Infrastructure
    - ***Storage:*** Deploy a scalable storage architecture with a capacity of 15-20 petabytes, utilizing high-performance, bulk, and archival storage solutions to accommodate the large volume of imaging data.
    - ***Compute and GPU Infrastructure:*** Establish a GPU compute facility with an initial setup of 50-100 GPU nodes, leveraging high-end GPUs for intensive computational tasks and AI/ML model training. This GPU cluster at NYP will complement the existing HPC cluster resources at Columbia and Cornell.
    - **Networking:** Implement high-bandwidth (100 GbE or faster) and low-latency networking solutions to support efficient data transfer and processing between storage, compute resources, the Flywheel platform, and the HPC clusters at Columbia and Cornell.
    - ***Cloud Availability:*** Azure Cloud Infrastructure (Either through flywheel or through NYPH)
    - ***Storage:*** Utilize NYPs Azure Blob Storage for cost-effective, scalable storage of imaging data, ensuring seamless integration with the Flywheel platform and in-house resources.
    - ***Compute:*** Leverage Azure Virtual Machines (VMs) and Azure Kubernetes Service [‘p(AKS) for additional compute resources, allowing for flexible scaling based on research demands.

    - ***Networking:*** Ensure secure and high-performance connectivity between Azure Cloud resources and the in-house infrastructure using Azure ExpressRoute or VPN gateways.
- Software Infrastructure
    - ***Flywheel Discovery:*** Implement Flywheel Discovery as the common backend for imaging metadata indexing and curation across all NYPH hospitals, leveraging its features for data organization and collaboration.
    - ***Flywheel Core:*** Deploy independent instances of Flywheel Core at Columbia and Cornell for research and development, allowing for customized workflows and integration with analysis tools
    - ***Development Tools and other software:*** Utilize a mix of subscription-based and open-source IDEs (e.g., JetBrains suite, VS Code) for code development, with a focus on flexible licensing to accommodate diverse development needs.
    - ***Version Control:*** Implement enterprise versions of GitHub, GitLab, or Bitbucket for enhanced collaboration, security, and CI/CD integration.
    - ***Containerization:*** Adopt Docker for application containerization, simplifying the integration of post-processing plugins as Flywheel Gears.
    - ***AI and Data Analysis:*** Incorporate open-source libraries (TensorFlow, PyTorch) for AI/ML development, supplemented by commercial platforms where necessary. Use a combination of open-source and commercial data analysis tools, optimizing for cost-efficiency and functionality.

4.1.8.1 Implementation Strategy

1. **Infrastructure Setup**: Establish foundational in-house and Azure Cloud storage and compute resources, focusing on scalability, performance, and security. Implement core software tools and platforms, ensuring compatibility and integration across the infrastructure.

2. ***Flywheel Discovery Deployment***: Deploy Flywheel Discovery as the common backend for all NYPH hospitals, configuring it to seamlessly integrate with in-house and Azure Cloud resources.

3. ***Flywheel Core Deployment:*** Set up instances of Flywheel Core at Columbia and Cornell, tailored to their specific research and development needs, with individual IRBs for collaboration on combined data.

4. ***Development Enablement:*** Set up advanced development tools, version control systems, and AI/ML libraries to facilitate algorithm development and research projects.

5. ***Optimization:*** Scale the GPU compute facility, storage resources, Azure Cloud infrastructure, and HPC utilization based on research demand and data

growth. Continuously evaluate and integrate additional software tools to enhance research capabilities and infrastructure efficiency.

This blueprint plans to leverage the strengths of in-house resources, Azure Cloud infrastructure, and the existing HPC clusters at Columbia and Cornell to create a powerful, scalable, and secure environment for managing and analyzing the vast amount of imaging data generated across NYPH, Columbia, and Cornell. By integrating Flywheel Discovery and Core, along with a robust software infrastructure will enable researchers to collaborate effectively, develop cutting-edge AI/ML algorithms, and drive advancements in medical imaging research at NYPH.

4.1.9 Outstanding Issues

There are substantial barriers in culture and infrastructure that need to be catalyzed. Governance for efficient, quick, and secure access to large scale data from the health system for medical research is a major barrier. Also having an efficient process in place for secure yet quick and easy to pursue process for cross-institutional collaborations. The current IRB, regulatory, and governance aspects are way bigger barriers than IT.

4.2 Dr. Michal Levo, Gene Regulation and Genome Organization

Content in this section authored by Dr. Michal Levo, Columbia University Irving Medical Center (CUIMC)

4.2.1 Use Case Summary

Dr. Levo is an assistant professor in the Department of Biochemistry and Molecular Biophysics, and the Program for Mathematical Genomics, at Columbia University Medical Center. She gained an interdisciplinary education with a dual undergraduate degree in Computer Science and Life sciences. During her PhD, in computational Biology, she employed high-throughput sequencing-based approaches to study how regulatory instructions are encoded within individual regulatory DNA sequences. She then perused a postdoc in Princeton University employing live imaging to quantify endogenous transcription dynamics during development and dissect their determinants in complex genomes.

As our mapping of eukaryotic genomes improved in the last few decades, we learned that genomes are replete with regulatory sequences guiding the spatiotemporal activation of genes, that in turn determines cell identity. Developmental disorders and diseases are often associated with disruptions to this regulatory code. Furthermore, we now know that the number of regulatory DNA elements exceeds the number of genes, and these elements can reside far away along the genome from their target gene. What are the mechanisms that mediate gene regulation within this genomic labyrinth remains an open question. The Levo lab combines several experimental modalities (from genomic assays to live sub-cellular imaging), with computational analysis and modeling, to study the dynamic, multilayered control of gene expression in the context of genome organization, particularly in the service of differentiation and development. Studies focus on perturbation-based approaches and quantitative measurements to facilitate causal insights and inform predictive models.

4.2.2 Collaboration Space

Dr. Levo expects collaboration with researchers at Princeton university, primarily working on the university cluster - hence a need to easily transfer large volume of data from that cluster to a CUIMC endpoint. There will be other collaborations with researchers in universities in the west coast well as Europe (e.g., at EMBL Germany). There will also be collaborations within Columbia - with researchers in department and in other divisions at the medical campus and on the main campus as well. These will necessitate transfer of large datasets and shared access to folders of raw data or analysis.

4.2.3 Instruments & Facilities

Dr. Levo's work will include genomic datasets - from high-throughput sequencing at Columbia and at external companies. The work further includes dataset on the order of tens of GB produced daily on a confocal microscope at the lab (Black building).

4.2.4 Data Narrative

Dr. Levo will focus on research that involves the in-house production of imaging data. The lab conducts continuous imaging sessions of live and fixed samples that produce files on the order of 1-10GB or 20-50 GB. At peak production – it will produce between 5 to 50GB files (depending on the experiment and consequently on data structure size) in a 4-hour imaging session and 2-3 sessions can take place during the day (and for fixed samples also over night and on weekends). In one modality of the microscope (super resolution Airyscan) the file size is larger, a 4h imaging session can yield almost 200GB files, but the data can be streamed directly to an offline computer on the same floor, via a 10Gbps internet connection, where an initial step of analysis is conducted reducing data size back to ~50GB file range.

The raw files need to be stored (most likely will be done on cold storage in AWS - maybe even via tape transfer), but only the "smaller" files are used for subsequent analysis. These files need to migrate to a cluster or to AWS for subsequent analysis. Analysis will be done by multiple users (e.g., students in the lab) running both parallel jobs of established pipelines or developing pipelines - necessitating interactive nodes.

Software like Matlab, R, Python, imajeJ/fiji and associated packages are routinely used. After initial imaging analysis, to extract key features from the data, post analysis is based on scripts per projects. These aim to plot various aspect of the data, conduct comparisons, overlay different data types and produce figures for manuscripts. In some projects machine learning algorithms are necessary for data analysis and simulations are carried out.

4.2.4.1 Data Volume & Frequency Analysis

Currently the Levo Lab produces GB scales of on an hourly basis.

4.2.4.2 Data Sensitivity

There are no sensitive aspects to the data used in the Levo Lab.

4.2.4.3 Future Data Volume & Frequency Analysis

The Levo Lab anticipates producing GB scales of data on an hourly basis in the future.

4.2.5 Technology Support

4.2.5.1 Software Infrastructure

Standard package for genomic analysis is used. In addition, we carry out image analysis, necessitating package and custom written code on Matlab, Python, and R.

4.2.5.2 Network Infrastructure

The Levo Lab is new and is currently trying to set up connectivity. The lab already encountered issues in physically connecting two computers at extreme end of the floor with a 10Gbps cat7 cables. The lab will need connectivity to clusters on campus (e.g., the current C2B2 cluster) and to AWS.

4.2.5.3 Computation and Storage Infrastructure

The Levo Lab is new and is currently trying to set up storage. As data needs to be backed up, the lab will strive to rely on clusters and HPC services of the university and if none are available with AWS (though cost issues are problematic).

Other universities currently offer superior storage and HPC services, there is largely no payment or reasonable fee for all-encompassing solutions in Princeton, NYU, UChicago, MSK, Jon Hopkins, etc. all offering suitable service for research labs with similar needs to the Levo Lab.

The services offered by the C2B2 cluster can address the Levo Lab needs, but the pricing (for storage and analysis) rules this out as a solution.

4.2.5.4 Data Transfer Capabilities

The Levo lab has experimented with Globus but has run into issues in getting it to work reliability and consistently with collaborators.

4.2.6 Internal & External Funding Sources

Startup package, Columbia-internal awards.

4.2.7 Resource Constraints

Insufficient data transfer performance, insufficient storage system space, difficulty finding or accessing data in community data repositories and ease of onboarding lab members on an interactive and parallel analysis scheme.

4.2.8 Ideal Data Architecture

No answer was provided to this section. Current needs are still being understood for the Levo Lab.

4.2.9 Outstanding Issues

Dr. Levo is currently trying to transfer data from Princeton to Columbia. This consists of ~15TB that needs to be hot and accessible for analysis, and ~60TB of cold storage data. The Levo lab is attempting to do this via Globus from the cluster at Princeton to an S3 bucket at AWS and had just recently succeeded to transfer most of the data using Globus.

Dr. Levo also needs to be able to transfer ~20GB files from a collaborator.

4.3 Dr. R Graham Barr, Respiratory Epidemiology

Content in this section authored by Dr. R Graham Barr, Columbia University Irving Medical Center (CUIMC)

4.3.1 Use Case Summary

One area of my research pertains to the epidemiology of lung structure using large-scale analysis of imaging data of the lungs, mostly using research computed tomography (CT) scans but also cardiopulmonary magnetic resonance imaging (MRI). The work integrates in environmental and genetic risk to define risk, subphenotype disease, and determine which measures of lung structure are prognostically important.

Most of the work in multicenter and multidisciplinary involving medicine, epidemiology, radiology, biostatistics, engineering, physics, etc.

4.3.2 Collaboration Space

We collaborate with about 40 institutions in the US, plus a few in Canada, the UK, and Germany. We generate data as part of multicenter studies, QC and read it via Reading Centers, and share it with the investigative group and more generally via NIH/NHLBI platforms.

4.3.3 Instruments & Facilities

Data sharing is complex and varies by project. The CT work is housed in BME/Mudd Building on local servers. The MRI work is more diffuse and less organized. We are frequently transferring large amounts of data from the Univ of Iowa via internet 2, which works for BME but not in my local office space. MRIs are usually transferred using a WebPAX housed at JHU.

4.3.4 Data Narrative

We have found over the years that local data ownership and systems are much more efficient for custom-built deep-learning algorithms than cloud-based options and note that NIH/NHLBI has been trying to develop a cloud-based system for image analysis for almost a decade without full success. That said, a cloud-based research PACS with local and remote (i.e., other institution) login would potentially work.

4.3.4.1 Data Volume & Frequency Analysis

Currently the Barr Lab produces TB scales, but the frequency is highly variable.

4.3.4.2 Data Sensitivity

Some of the data that the Barr Lab produces can be considered sensitive, but there are appropriate controls in place for dealing with this data. All images are supposed to be deidentified but 1) deidentification has failed in the past, 2) imaging is to some degree identifiable; therefore, we treat it as such.

4.3.4.3 Future Data Volume & Frequency Analysis

The Barr Lab anticipates producing TB scales of data, with a similar variable frequency.

4.3.5 Technology Support

4.3.5.1 Software Infrastructure

Most of the DL work is performed by my BME collaborators using custom scripts in Python, C++, etc.

4.3.5.2 Network Infrastructure

BME/Mudd connectivity is now excellent, unlikely several years ago. But image transfer in many local parts of the Medical Center remains slow such that encrypted hard drives remain a good option.

4.3.5.3 Computation and Storage Infrastructure

4.3.5.3.1 COLUMBIA UNIVERSITY MEDICAL CENTER (CUMC)

Situated on a 20-acre campus in Northern Manhattan and accounting for roughly half of Columbia University's nearly $3 billion annual budget, Columbia University Medical Center (CUMC) provides global leadership in scientific research, health and medical education, and patient care.

CUMC's major teaching hospital affiliates are NewYork-Presbyterian Hospital and the New York State Psychiatric Institute, both of which share the CUMC campus. The medical center also has academic affiliations with Bassett Healthcare, in Cooperstown, NY; the Isabella Geriatric Center, in New York City; and Arnot Ogden Medical Center, in Elmira, NY; and Stamford Hospital in Stamford, CT. Columbia's faculty practice is ColumbiaDoctors.

4.3.5.3.2 Columbia University College of Physicians & Surgeons (P&S)

The College of Physicians & Surgeons (P&S), founded in 1767, was the first American school to award the MD degree in 1770, and its legacy of innovation continues. Major curriculum changes implemented in 2009 reflect the changing practice of medicine in the 21st century. The new Bassett Track trains students for primary care work in rural environments. P&S is one of the most research-intensive medical schools, and its students' MCAT scores and grade-point averages are among the highest in the country. Furthermore, students seeking their biomedical science doctoral degree have access to state-of-the-art resources, facilities, and leading faculty in a variety of fields.

4.3.5.3.3 Dr. Barr's Laboratory

Laboratory: Dedicated laboratory space for the project is located in 200 square feet of renovated, dedicated laboratory space in the Division of General Medicine.

Clinical: The project will utilize already existing infrastructure established through the MESA and SPIROMICS studies. This includes 250 square feet of renovated, committed clinical examination space in the Columbia University Clinical and Translational Science Award (CTSA)/Irving Institute. This space houses an interview room, pulmonary function testing equipment and a sputum induction room. Additional space in the CTSA includes a waiting area, consent room, phlebotomy room, interview rooms and an established six-minute walk test course. The CTSA is located one floor above Dr. Barr and the project staff's offices in the Division of General Medicine.

Computer: There are approximately 350 desktop and laptop computers in the Division of General Medicine, one for each faculty member, research assistant and administrative staff. These business computers are connected to the Columbia University network and by high-speed access to the internet via both cable and Wi-Fi connections. These include high-performance workstations for intensive image processing (e.g., Dell Precision 7920 Tower with Intel Xeon Gold 6130 2.1GHz, 3 .7GHz Turbo, 16C, 10.4GT/s 3UP I, 22M Cache, HT (125W) DDR4-2 666 2nd), 64 bits, 32 GB RAM, 2TB HD) running Matlab, Ensight, Mimics, Osirix, TreeAge, Stata, Python, R, Rstudio.

The Division of General Medicine leases 6 TBs from a virtualized File Server. The virtual sever consists of discs scalable up to 2.4 PB. Users are authenticated through Active Directory. Computers outside the institution will need to login through VPN to the access the university network. Each user still needs to authenticate through Active Directory to access the file storage. Also, each user can be granted only specific directories to enhanced security. The virtual server is a certified system under CUMC IT Information Security. CUMC IT Information Security has adapted the CSF to perform application assessments to enhance our security compliance posture for HIPAA HITECH, PCI/DSS, NIST and SSN protection. There are at least 3 backup copies made throughout the day and a final one done daily at night. Backups are kept on tape up to a year.

The Division of General Medicine leases two virtualized Windows 2019 R2 Servers to host Filemaker Databases. One server is dedicated to web-publishing while the other is dedicated to database functionality for desktop, web clients and mobile clients. The basic specifications of the two virtual servers are: 4 Processors per Server @ 2.70 Ghz, each with 16 GB RAM, x64-based processor and 150 GB of space provided between OS and application. Database Users are authenticated through Active Directory. Computers outside the institution will need to login through VPN to the access the Database. The database can be served in multiple platforms: Windows, Macs, Web, thin Clients, ipads, iphones, Android phones and tablets. All communication between client and server is authenticated using custom SSL Certificates (Comodo Elite SSL Certicates). Comodo Elite SSL Certificates provide 256 bits standard SSL encryption, and This Certificate is inherently recognized by 99.9% of the current Internet population browsers. These virtualized application servers are certified under CUMC IT Information Security. There at least 3 backup copies made throughout the day and a final one done daily at night. Backups are kept on tape up to a year.

Qualtrics Surveys: Qualtrics is a web-based survey software tool available for use by all faculty, and staff. Qualtrics can be used to capture survey results from a publicly available survey, or from potential participants or eligible participants who are specifically given access to a survey. Qualtrics is a certified system under CUMC IT Information Security. Various tools are available to create interfaces between Qualtric Surveys and databases (Filemaker, SAS) to have the data available as soon as participants entered the data. The data is secured and encrypted as well. There are tools that are available to mobile platform as well. It is possible to have surveys sent via SMS text or Mobile Browsing Data or Secured Email.

Microsoft A5 License: CUMC obtains this enterprise license. Each user is supplied with the MS Office 365 products, including 1 TB of OneDrive, 100GB of Outlook, Teams, and SSO.

DUO MFA: All major applications are connected to DUO MFA. This includes VPN, MS Office 365, and Adobe acrobat Creative Cloud Suite, Box, Zoom, etc.

Box Application: to communicate with collaborators outside of the institution, Box is offered to CUMC users.

The Division licenses various software like SPSS, SAS, Adobe Acrobat Professional, and MATLAB.

Office: Office space for Dr. Barr and the project staff are in dedicated space in the Division of General Medicine. The Division occupies more than 8,000 square feet of renovated office space for dry laboratory research.

Other: The Division of General Medicine has two full-time Information Technology specialists, a database programmer, a research database manager, five grants' managers, and two personnel/office managers.

Major Equipment: Pulmonary function equipment includes a rolling barrel Sensormetics/OMI spirometer, an Inspire spirometry system, multiple ndd EasyOne Diagnostic spirometers, two ndd EasyOne Pro DLCO system, a Hans Rudolph DLCO analyzer, and a sputum induction system and hood.

Available equipment in General Medicine laboratories include: one microfuge, two large industrial centrifuges, two -80C freezer, two -20C freezer, two laboratory refrigerator, cone and platelet analyzer, optical aggregometry machine, optical microscope, ambulatory blood pressure machines, Finapress machine, ELISA plate shaker, and ELISA reader.

4.3.5.3.4 The Irving Institute for Clinical and Translational Research at Columbia University

The mission of Columbia University Medical Center's Clinical and Translational Science Award (CTSA) is to transform the culture of research to hasten the discovery and implementation of new treatments and prevention strategies. CUMC's first step towards reaching this goal was the establishment of the Irving Institute for Clinical and Translational Research, the academic home for patient-oriented research at Columbia. The Irving Institute faculty includes some of CUMC's most senior researchers who provide leadership and serve as mentors for junior faculty, fellows, and trainees. The resources provided by the Irving Institute offer support in biomedical informatics, study design and biostatistics, bioethics, regulatory issues, core laboratory facilities and a fully staffed Clinical Research Center for investigators across campus.

The Irving Institute awards over $1 million in pilot funding each year to Columbia University faculty. The Pilot and Collaborative Studies Resource (PCSR) is central to achieving the goals of the Irving Institute by providing incentives to both young clinical and translational investigators, as they obtain pilot data prior to submitting funding applications, and more senior investigators who may not otherwise engage in multi- and interdisciplinary research. PCSR aims to establish a coordinated structure and collaborative environment within which multi- and interdisciplinary clinical and translational research may flourish by expanding and optimizing the utilization of outstanding new and existing resources on CUMC, training a new generation of multidisciplinary clinical and translational investigators, enhancing recruitment of exceptional clinical and translational investigators to pre-clinical and clinical departments at CUMC, and creating incentives for mentoring junior faculty and outreach from one department or school or campus to other interested researchers.

The Irving Institute devotes significant institutional resources to activities related to Clinical and Translational research, including staff, clinical research, and office space.

Office: The Irving Institute's headquarters occupies the entire 10th floor of two connected buildings, the Presbyterian Hospital, and the Harkness Pavilion. The total space occupied is approximately 25,000 square feet, including a spacious education and training hub that includes a large dividable classroom with a capacity of 72 and a conference room with a capacity of ten. The administrative center occupies 1,211 square feet at the center of the overall space. Immediately adjacent, are both the Adult Outpatient Unit (2,450 square feet) and the Adult Inpatient Unit (9,300 square feet). Across the hall are 690 square feet of space dedicated to the Bionutrition Unit. Also included is a cluster of office space totaling 1,000 square feet which houses additional Irving Institute staff, faculty, and collaborators. A 400-square-foot conference room and 5,000 square feet of laboratory and office space used for the Biomarkers Core round out our state-of-the-art location. In an adjacent building is the 2000 square foot Pediatric Outpatient Unit.

4.3.5.3.5 Irving Institute CTSA Information Technology

CTSA IT hosts multiple systems that provide services to the CUMC research community. Each runs on one of seven Windows 2008, Windows 2012, and Solaris servers. One of these is a data management system that supports over 100 active clinical research studies. CTSA IT provides desktop support, network management, and consulting services; it works closely with other IT groups at CUMC to coordinate and standardize computer security. The CTSA IT programming team provides expert advice on the design, development, and deployment of data management systems and other scientific and administrative applications.

4.3.5.3.6 Columbia University Medical Center Information Technology

Information and Communication Technology Infrastructure and Services at CUMC are organized under the umbrella of Columbia University Medical Center Information Technology. (http://cumc.columbia.edu/it/about.html)

The CORE Resources Group is a joint group between the New York Presbyterian Hospital and Columbia University Medical School. The name represents the "central" or "core" role we play in the day-to-day operations of electronic data communications for the institutions we serve. The group is composed of both Hospital and University employees. Our responsibilities are primarily the design, implementation and maintenance of the New York Presbyterian Hospital and Columbia University Medical School Health Sciences Campus network infrastructures. Our network provides secure transport for high-speed data, IP Telephony, and video traffic. High speed network access is available in all the CUMC campus buildings. Portions of the CUMC campus have 802.11g/n wireless available to our user community. We maintain and operate a privately owned 40-Gbps Wide-Area-Network (WAN) with redundant 10-Gbps fiber optic links between 12 major NYP and CUMC campuses. Internet service for the CUMC campus is provided by the Columbia University Morningside campus via redundant 10-Gbps fiber optic links between the Morningside and CUMC campuses. Remote offices and practices connect back to our main campuses via high speed, dedicated links. Other services we design, implement and support are DNS, DHCP, IP Management, and remote connectivity via VPN. Our network is enabled for multicast routing.

The CUMC IT Server Support team provides premium file, print, application, and data backup services, as well as many other resources for a wide range of technical demands. Our group of experienced IT professionals has supported the CUMC campus for close to a decade. We consistently strive to provide Faculty, Staff and Students with cutting edge technology solutions and reliable, secure data storage. The Server Support group provides resources for many departments at CUMC, including Health Sciences Central Administration, DBMI, and the Ideatel project. This includes management of domains and logon IDs for departmental computers, network drive space, the CUMC Exchange mail and calendar server, Network-Attached Storage (NAS) systems for easily accessible load-balancing, fault-tolerant storage, Virtual Tape Libraries for quick and reliable access to archival data, and much more. Our servers are housed in state-of-the-art storage facilities that include

24-hour temperature and moisture monitoring to help prevent hardware failure, and our backup services include offsite storage to meet best practices for disaster recovery plans.

The Information Security Department is a joint group between the NewYork-Presbyterian Hospital and Columbia University Medical Center. We're comprised of staff from both NYP and CUMC who are responsible for administration of security policies for the technical infrastructure on campus. We're also here to guide technical groups in compliance with HIPAA regulations. Networking at the lowest layers interconnects most of the computing devices in a medical center today. The flow of information among these systems is complex due to the variety of transport mechanisms and the volume of such information. The iterative process of simplifying the exchange methods, generalizing as well as making them more robust, is an engineering challenge. The creation of real transfer of patient care information using open systems and building standards to accomplish it are some of our active work areas.

Distributed computing is a reality at CUMC; we are currently learning how to build systems that may be monitored and managed to improve availability and reliability. The semantic correctness of data has direct impacts on quality of care and reduction of costs. We have begun work on techniques to improve secured but easier access to health care information by providers and researchers. This work has direct relevance in creating nation-wide repositories of clinical information.

4.3.5.3.7 Columbia University Shared and Core Resources

The Medical Center is the site of Core Labs and facilities with expert staff available for consult and assistance in a wide range of specialties. Columbia University Medical Center has more than 60 state-of-the-art shared research facilities physically housed in and administered by its departments, centers, and institutes. Clinical research faculty, basic scientists, and students all benefit from the shared access to and cost of these facilities. These resources offer the highest-quality scientific technology to the community. Many of these facilities also offer education and training, as needed.

Biomarkers Laboratory: The Biomarkers Laboratory is a shared Facility of the Herbert Irving Comprehensive Cancer Center and Columbia's NIEHS Center for Environmental Health in Northern Manhattan. The Core provides initial consultation on all aspects of sample collection and processing for specific studies. Biological samples received are coded to maintain confidentiality using preprinted bar code labels provided to investigators. The Core also processes and stores urine samples. Samples are stored in -80oC or liquid nitrogen freezers, all connected to a Sensaphone alarm system which calls out any freezer or room temperature fluctuations. Multiple laboratory staff are on 24/7 call to address issues as they arise. DNA is extracted using various protocols and quality control involves the determination of the ratio of the absorbances at 260 and 280 nm. The Core also has the capacity to carry out whole genome amplified (WGA) DNA. Candidate SNP genotyping is carried out using Applied Biosystems Taqman kits in 384 or 96 well plates.

A web-based database is used for the inventory of stored samples in conjunction with the bar code reading system. Each submitted sample is identified by a unique sample ID. All samples are labeled with a bar code representation of the sample ID. The database allows linkage of the sample inventory with the information available from questionnaires, surveys, etc. through the unique sample Facility Core number. Information in the database includes the study identifiers, sample ID, aliquot type, volume, location (freezer, shelf, rack, box, and position in box), date received, date processed and the technician who performed the processing. The database also records all use and shipment of specimens, whether for internal use or by outside institutions and researchers. The database system employs a number of internal checks to ensure the integrity of the inventory data. In addition to the scanning of all sample IDs, the system uses an interface created in Microsoft Access, which provides a feature-rich interface for the user (dynamic querying and searches, automatic entry of default values, automation of routine tasks). Data are stored permanently in an MS SQL Server 2000 database, which enforces quality constraints and referential integrity (e.g., ensuring that all samples can be properly linked to a study and individual submission). Access to the data is via an encrypted HTTP (internet) connection, and access is provided for authorized users with valid passwords only. A separate web-based database system stores genotyping data including the methods used and test results.

4.3.5.3.8 Columbia Institutional Review Board (IRB)

The Columbia University Institutional Review Board (IRB) offers training to supplement research compliance and human research protection requirements, including: monthly investigator meetings; quarterly IRB 101 informational sessions; annual Institutional Review Board educational conferences; and web-based or in-person training on conducting research with minors, conducting research with animals. Weekly open consultation hours are also offered by the IRB. Investigators may meet with IRB professionals to discuss regulatory, policy, and procedural questions. Mandatory, web-based training on good clinical practices, human subjects' protection, HIPAA policies, and research with minors (if applicable) is required for any individual involved with research at Columbia University. Offices are conveniently located on both the Morningside Campus as well as Columbia University Medical Center (http://www.cumc.columbia.edu/dept/irb/).

4.3.5.3.9 CUMC IRB Liaison Service

Jointly supported by the Human Research Protection Office (HRPO) and the Irving Institute, this free service is available to CUMC researchers seeking assistance with understanding and addressing IRB requirements and requests. The IRB Liaison will serve as a link between the IRB and CUMC investigators who have submitted a protocol for review by one of the Columbia University Medical Center IRBs. The primary objective for this service is to improve the quality and efficiency of human subject research protocol submissions and responses to IRB requests. The IRB

Liaison will provide consultation in preparing protocols to be compliant with IRB requirements. In addition, the IRB Liaison will provide support to investigators for responding to IRB reviews of research protocols, explanation of IRB requests and assistance in providing appropriate responses and/or implementing requested changes. Consultations from the IRB Liaison are in addition to existing consultation services provided by IRB staff.

4.3.5.3.10 THE DAVID A. GARDNER PET IMAGING RESEARCH CENTER

The David A. Gardner PET Imaging Research Center at Columbia University Medical Center is a comprehensive imaging facility that includes 2 cyclotrons, an FDA compliant production radiochemistry and radio pharmacy, and 4 scanners. In addition, there is a fully equipped cold chemistry space capable of conjugation chemistry and radiochemistry, including halides (I-124) and radiometals (Zr-89).

A research radiochemistry laboratory is located on first floor of the building and has 3 hot cells with Remote G Manipulators. This laboratory is used for the development of new radiotracers or validating existing tracers to establish synthetic procedures to move them to cGMP laboratory. This lab has been equipped with FXMEI, FX-FN, and CN modules.

There are Quality Control (QC) and Microbiology Lab spaces established for radiotracer QA/ QC analyses. The QC space is well equipped with analytical HPLC quality control stations, gas chromatography for volatile organic component testing in the final product, Radio-TLC, MCA analyzer, pH meter. The three Agilent HPLC QC systems consist of one 1150 and two 1200 systems outfitted multi-wavelength UV detectors and a FlowRAM sodium iodide radioactivity detectors. The outputs from all the detectors are processed through three independent LabLogic Laura workstations. The entire setup provides fast and reliable results of chemical, radiochemical and specific activity. The Microbiology laboratory is well equipped with a laminar clean bench for sterile inoculations, as well as incubators for sterility verification, and two Charles River's PTSpyrogen testing devices.

Metabolite samples in sealed isolated tubes are sent back to metabolite lab adjacent to PET camera area. The metabolite lab is equipped with centrifuges, an HPLC setup incorporating both UV and LabLogic Posi-RAM radioactivity detector and automated gamma counters which allow for determinations of metabolite formation over time.

The facility also has pre-clinical capabilities operated by a qualified veterinary technician. The space is equipped with a Siemens Inveon small animal PET scanner. The space also has rodent housing capabilities and a biology lab outfitted with a dissection microscope. The gamma counters also support small animal organ biodistribution analyses following dissection.

The David A. Gardner PET Imaging Research Center has one Siemens 64 Slice mCT-SBiograph

PET/CT scanner, and two Siemens 64 slice Biograph mCT Flow™ capable of performing PET, CT and PET/CT imaging of humans and large animals (ex: non-primates). The research radiochemistry space has distinct areas for cold chemistry including precursor development, conjugation chemistry (including a metal-free workspace for radiometal conjugation and chelation), and halide chemistry (with fully compliant fume hoods and laminar hoods).

4.3.5.3.11 THE DEPARTMENT OF RADIOLOGY

In 2021, the Department of Radiology performed over 534,892 imaging procedures across a variety of locations, including NYP Columbia, NY Children's Hospital, NYP Allen Hospital, and offsite Columbia Doctors Radiology satellite locations. The department offers a world class residency program in Radiology and a CAMPEP-accredited imaging medical physics residency program. With over 100 faculty and more than 60 trainees, the Department is among the top academic radiology departments in the country.

4.3.5.3.12 Imaging Infrastructure and Clinical Facilities

The Department of Radiology is organized into seven specialized clinical divisions: Abdomen and Chest, Interventional Radiology, Nuclear Medicine, Musculoskeletal, Pediatrics, Breast Imaging, and Medical Physics. The clinical imaging infrastructure is expansive, featuring 13 MRI units, 17 CT scanners, 4 PET/CTs, 3 SPECT/CTs, 40 odd ultrasound units and 150 plus x-ray tubes. All of these are integrated into a fully functional PACS system running on a medical-grade network. The department also houses a state-of-the-art 3D lab, the Center for Advanced Image Processing and Analysis (CAIPA), which optimizes radiologist workflow by providing routine post-processing for all clinical care.

4.3.5.3.13 Radiology – Division of Medical Physics

The Medical Physics Division consists of eight medical physicists. The division maintains strong collaborative ties with faculty and students from the departments of Biomedical Engineering (BME) and Applied Physics and Applied Mathematics (APAM). These collaborations primarily focus on translational research projects, including the application of machine learning and deep learning techniques for MR image reconstruction and disease classification. Dr. Jambawalikar not only leads the Medical Physics Division but also directs the Medical Imaging Physics Research Lab. This lab collaborates extensively with Radiology faculty and residents, as well as with faculty and students from BME and APAM. The team has been involved in a variety of medical imaging projects, particularly those that leverage deep learning for MR image reconstruction and disease classification. Some members of the team hold joint appointments with the BME and APAM departments at Columbia University. Graduate students from these departments also contribute to research projects in collaboration with Radiology faculty and residents on the CUMC campus.

4.3.5.3.14 Deep Learning Medical Physics Computational Resources

The Medical Physics Lab is a state-of-the-art facility tailored for advanced AI and medical research. The lab is proficient in various machine learning and deep learning frameworks like TensorFlow, PyTorch, and Keras, and is fully integrated with hospital research systems, including PACS. The lab is equipped with:

1. 7 high-performance PCs, each with either an Intel Xeon E3-1220 Quad Core @ 3.0GHz or an Intel Xeon E5-1620 Quad Core @ 3.7GHz, at least 64 GB of memory, and 2TB of storage.

2. Departmental Nvidia GPU cluster with 8 A100 GPUs each with 80 GB of GPU memory.

3. The lab boasts 13 GPU cards and an NVIDIA DGX station, as well as a multi-CPU system featuring a 36-core Xeon processor, 256 GB RAM, and 50 TB of RAID10 storage, all operating on an Ubuntu Linux platform.

4. 2 servers equipped with 2 NVIDIA GeForce RTX GPUs with 24GB GPU RAM with NVlink, specialized for advanced AI research.

5. A dedicated terminal for hospital PACS (Philips Vue Pacs).

6. A DICOM image receiving server with a permanent license for eFilm Workstation® 4.2.2.

7. A 125 TB shared physics drive.

8. A petabyte of CUIMC/NYP box storage for data sharing, storage and collecting data from sites

9. A server with QuPath installed for pathology image viewing, annotation, AI analysis and storage.

10. Radiology REDCap server for collecting research data /information.

All servers and the shared drive have multi-location backup systems. The lab is connected via the hospital's 1TB Ethernet network. Each high-performance PC is equipped with dual 23-inch widescreen monitors, and four of these PCs also feature NVIDIA GeForce GTX 1080 GPUs.

The initial phases of AI studies, such as model selection and hyperparameter tuning, are conducted on the 2 NVIDIA RTX systems. The later stages, which involve large-scale image data training, are executed on the Nvidia DGX system with 4 V100 GPUs or the Departmental resource of 8xA100 GPUS. The lab has also developed a web-based imaging system, hosted on the research PACS server, utilizing open-source tools like Weasis and DCM4CHEE. This system offers a comprehensive range of visualization functions, advanced lesion and organ segmentation tools, manual editing capabilities, and an efficient workflow for therapy response assessment. All lesion and organ data are securely stored in a dedicated relational database server.

Clinically acquired DICOM images can be transferred from the hospital PACS to our DICOM receiving server. A custom de-identification program then removes all patient-identifiable information (PHI) before storing the de-identified images on the shared drive. The lab provides access to a variety of commercial and open-source software packages, such as Matlab and 3D Slicer, as well as proprietary image processing and visualization tools for a wide range of research projects.

4.3.5.3.15 Image Analysis Resources

The department maintains a 32-node Rocks cluster for in-house image processing and statistical analysis. The CAIPA lab is equipped with server and workstation solutions for advanced image processing and analysis, including various neuroimaging packages and image registration software.

4.3.5.3.16 Radiology Information Technology

The department's IT infrastructure is robust, featuring a 165TB Philips PACS system, 57 reading workstations, and an EPIC RIS system. The department also boasts a massive data storage infrastructure with a capacity of 2 petabytes. The Nuance PowerScribe 360 dictation system is integrated with PACS, streamlining the report generation process.

4.3.5.3.17 Columbia Radiology MRI Center

The lower level of the Neurological Institute Building on the Columbia-Presbyterian Campus houses the Columbia Radiology MRI Center. The Columbia University MRI Center at the Neurological Institute offers the Columbia research community access to state-of-the-art imaging on General Electric (GE) 3T Premier systems. There are currently four GE 3T Premier scanners with one dedicated system for research. Radiology MR physicists and onsite GE scientists provide support for research on these systems. In addition, as part of a MR partnership with GE Healthcare to support the clinical and research endeavors, GE has their Applied science (ASL) East laboratory (a team of 4 GE scientists) located onsite at the Radiology MRI Center. All 4 systems have GE research mode capabilities with multiple prototype GE research sequences installed on the scanners. Research prototype pulse sequences for Multiband DTI for Connectome, UTE (Ultrashort TE), QSM (quantitative susceptibility mapping), 3D PCVIPR for 4Dflow, EPI-MIX, MR Fingerprinting and multiple GE work in progress sequence are installed and available on the scanners for research purposes. All standard and advanced pulse sequence options and applications for; Neuro applications, Cardiac applications, Musculoskeletal applications, and Body applications are available on the scanners. All scanners are equipped to synchronize visual stimulus with a rear projection, audio stimulus and physiological data (EKG, peripheral pulse, respiration). Multiple stimulus response systems such as button response, trackball, and joystick is available at all scanners. Advanced post processing software for Image registration and functional image analysis and spectroscopy (LCModel analysis) is available through workstations and Linux CPU and GPU server supported by Radiology IT. In addition, the radiology

Department has a 96 TB research data storage server at NI basement with redundant backup at PET center.

4.3.5.3.18 HEFFNER BIOMEDICAL IMAGING LABORATORY

The Heffner Biomedical Imaging Laboratory facility is located on the Morningside campus of Columbia University, in the Fu Foundation School of Engineering and Applied Science (SEAS), Seeley W. Mudd Building, 500 West 120th St, New York, NY 10027.

The Heffner Biomedical Imaging Laboratory occupies a space of 1,050 square feet on the Engineering / Morningside Heights Campus of Columbia University, located at 116th street and Broadway. The laboratory has desks for 8 researchers, and 2 postdoc / visiting scholars along with a meeting space for small group conferences with video / teleconferencing resources.

Offices: Prof. Laine has his primary office in the Biomedical Engineering Department on the Morningside campus, with support staff and other basic office resources. He also occupies a secondary office on the Columbia University Irving Medical Campus (CUIMC), with several carols for graduate students to work on the medical campus and meet / interact with clinicians and other basic scientist who work on the CUIMC campus during development of the proposed study.

Part of the Department of Biomedical Engineering's footprint on the medical campus, includes a newly renovated space located on the 5th floor of the Alianza Building. The building is located in the heart of the Columbia University medical school campus / New York Presbyterian Hospital and includes a large conference room (shared), offices for post-docs, and direct access via a HIPPA and PHI secure / compliant workstation to the New York-Presbyterian Hospital PACS systems.

The recruited pre-doctoral and/or post-doctoral fellow will also have a desk in the Heffner Biomedical Imaging Laboratory as well as on the medical campus.

4.3.5.3.19 UNIVERSITY OF BRITISH COLUMBIA

Dr. Hackett's laboratory is located at the Centre for Heart Lung Innovation (HLI) at the University of British Columbia, Vancouver, Canada. The HLI facility includes 50,000 sq ft of wet and dry laboratory space, including technology cores; 1) Molecular Phenotyping 2) Tissue Culture 3) Cellular Imaging and Biophysics 4) Preclinical Services 5) Cardiovascular and 6) James Hogg Lung Registry biobanks 7) Histology and 8) Information Technology. The cores are supported by technical staff, through the HLI annual infrastructure budget of $1.2M. Specifically, in the histology core for cutting tissue samples Dr. Hackett will use a microtome (Leica RM2145 Rotary Microtome), tissue embedding equipment (Leica EG 1140 Paraffin embedding system), cold plate (Leica EG 1140C), slide warming/flattening table (Leica HI 1220), water bath/tissue flattening bath (Leica HI 1210), and a tissue processor (Lecia TP1020).

Dr. Hackett's own laboratory space (3000 sq ft) includes a primary tissue culture room including a FlexCell bioreactor and Electric cell impedance sensing (ECIS) system, two C02 incubators (Thermo scientific/Napco 8000DH), two laminar flow hoods, liquid nitrogen storage tanks for cryopreserved cells with automated electronic alarms, and a fully equipped open space laboratory where the general biochemistry and molecular biology experiments are performed. The laboratory is supported by the IT core of the HLI and all data are stored and backed up daily on the HLI server with off-site data back-up for disaster recovery.

Dr. Hackett is also the Director of the Imaging mass spectrometry core based at the biomedical research center (BRC) within the UBC Flow Facility (https://flow.ubc.ca/). This newly acquired equipment (Hyperion-Helios, Fluidigm) was purchased in March 2022, with Canada Foundation Innovation funding ($1.2 M) for which Dr. Hackett was the principal applicant. The flow core has been using suspension mass spectrometry for the last 10 years and has a dedicated UBC antibody core for conjugating antibodies with metal isotypes (https://flow.ubc.ca/). To achieve high-quality data from the single-cell imaging of human samples, we will purchase tissue imaging validated antibodies from well-established suppliers (Fluidigm) and rely on the quality control standards of the antibody core that have been providing metal-tagged antibodies for the last 10 years.

The proposed project will also utilize the James Hogg Lung Registry (JHLR) for which Dr. Hackett has been the Director since 2014. The JHLR was established in 1977 and is now one of the largest lung tissue biobanks in the world, with over 40,000 samples from 3000 well-phenotyped patients, with different respiratory conditions including acute respiratory distress syndrome (ARDS), asthma, chronic obstructive pulmonary disease (COPD), cystic fibrosis, interstitial lung disease, and lung cancer. The majority of subjects have lung function, pathology, radiology, patient symptom questionnaire, and blood work from the time of sample acquisition. All of the human lung tissue samples will be accessed from the JHRC biobank. The JHRC contains a dedicated specimen preparation room with a pathology down-draft table, Nikon camera and specimen photography stand, and lung coring room with a fume hood and band saw for processing frozen human lungs. The lung registry has access to a research-dedicated high resolution computed tomography (HRCT) scanner for scanning frozen lung samples and 10 high capacity -80C freezers for storing samples with real-time temperature monitoring and backup power. These facilities will be used to process the new asthma donor lungs into the biobank.

4.3.5.3.20 JOHNS HOPKINS SCHOOL OF PUBLIC HEALTH (JHSPH)

JHSPH is the top-ranked school of public health in U.S. and the largest school of public health in the world, with nearly 900 full-time faculty members and over 3600 full-time and part-time students from more than 90 nations. It currently offers educational programs at the master's and doctoral levels in diverse areas ranging from molecular microbiology and immunology to health policy. JHSPH has become internationally recognized for both the excellence of its scholarship and its many contributions to the improvement of human health throughout the world.

JHSPH was first established in 1916 with a grant from the Rockefeller Foundation. The School's most recent mission statement indicates its central commitment to research and education: The Johns Hopkins Bloomberg School of Public Health is dedicated to the improvement of health for all people through the discovery, dissemination, and translation of knowledge, and the education of a diverse global community of research scientists, public health professionals, and others in positions to advance the public's health.

The School is structured around 10 academic departments. These include:

- Department of Biochemistry and Molecular Biology
- Department of Biostatistics
- Department of Environmental Health & Engineering
- Department of Epidemiology
- Department of Health, Behavior, and Society
- Department of Health Policy and Management
- Department of International Health
- Department of Mental Health
- Department of Molecular Microbiology and Immunology
- Department of Population, Family and Reproductive Health

The Department of Epidemiology at JHSPH is the oldest Department of Epidemiology in the world. The mission of the department is to improve the public's health by training epidemiologists and by advancing knowledge concerning causes and prevention of disease and promotion of health. The specific goals of the department are: (1) to provide the highest quality education in epidemiology and thus prepare the next generation of epidemiologists; (2) to advance the science of epidemiology by developing new methods and applications; (3) to use the methods of epidemiology to investigate the etiology of disease in human populations; (4) to use epidemiologic methods in evaluating the efficacy of preventive and therapeutic modalities and of new patterns of health care delivery; (5) to develop methodologies for translating epidemiologic research findings into clinical medicine; and (6) to develop approaches for applying the findings of epidemiologic research in the formulation of public policy and to participate in this formulation and the evaluation of the effects of such policy. The current Chair of the department is world-renowned HIV researcher Dr. David Celentano. Dr. Matsushita, Principal Investigator at Hopkins and Ms. Chen have primary appointments in the Department of Epidemiology at JHSPH.

The Welch Center for Prevention, Epidemiology, and Clinical Research: The Welch Center (Director: Lawrence Appel) is an inter-disciplinary academic unit with 31 core faculty and ~100 associate faculty members located in the 2024 East Monument Street Building, which is also the location of the Institute for Clinical and Translational Research (ICTR) and the Center on Aging and Health (COAH). The

Welch Center has ~16,000sqft, of contiguous space, and is only a 5 min walk from the Johns Hopkins School of Medicine and SPH. Dr. Matsushita is core welch center faculty and is regularly attending center activities. The Welch Center provides offices for faculty, research assistants, administrative staff, and trainees. Faculty members are organized into informal program clusters which allows for collaboration across research areas. The Welch Center has five conference rooms and a large number of carrels with computers and network access for trainees.

Welch Medical Library: The library is located on the Medical Campus and contains over 300,000 bound volumes, 2300 audiovisual programs and receives over 2800 biomedical periodicals. The library sponsors a number of computer software training programs, grant-writing courses and offers an extensive electronic journal collection. This is the largest medical library in the nation after the National Library of Medicine.

Computer: The ARIC investigators at Johns Hopkins are part of a large department of epidemiology within the School of Public Health, which maintains a state-of-the-art information system designed to carry out the computer-associated aspects of the data management and analysis in epidemiological studies. The unit is managed by experienced senior programmers, with additional part-time programmers available to work on specific projects. The manager and programmers may assist in or actually carry out these tasks for faculty-directed research projects. All computers are networked, and VPN and internet access protocols provide secure access from off-site locations to e-mail and internet files. A JHU network has extensive statistical and mathematical packages available, including Stata, SAS, SPSS, SPS, BMD, GLIM, MINITAB, and IMSL. All major programming languages are supported. A large library of programs specific to the needs of epidemiological research has been developed for data entry, editing, crosstabulations, linear and logistic regression, and other epidemiological analysis. Ad hoc programs are developed as required for special purposes.

Office: All faculty have an office including state-of-the-art computing, internet and e-mail programs as well as office and data analysis software as needed. Dr. Matsushita and many faculty and trainees have offices at 2024 E. Monument, 2-blocks away where they interact closely with faculty members in JHSPH and School of Medicine. A single e-mail server supports both offices allowing for transparent movement of information as Dr. Matsushita and others move across offices.

Other: A well-trained highly centralized unit supports administrative efforts in the Department of Epidemiology. The group has extensive experience managing contracts and grants from the NIH, other government agencies, and private agencies. The divisional based Research Administration Unit provides post-award support. An individual from that unit is named the contractual contact. Purchasing, accounts payable and research accounting are University-based. The University has extensive experience in managing the administrative aspects of projects such as the current proposal.

4.3.5.3.21 UNIVERSITY OF WASHINGTON

The UW is one of the world's preeminent public universities. UW is a multi-campus university in Seattle, Tacoma and Bothell, as well as a world-class academic medical center, and confers more than 12,000 bachelor's, master's, doctoral, and professional degrees annually. Ranked No. 8 on the U.S News & World Report's Best Global Universities rankings, the UW educates more than 54,000 students annually.

The UW School of Public Health, ranked among the top 10 public health schools in the US, is grounded in teaching, research, and service. For more than 40 years, our 10,000 graduates have gone on to transform communities, lead health organizations, and find solutions to emerging public health challenges. Major departments are Biostatistics, Environmental and Occupational Health Sciences, Epidemiology, Global Health, and Health Services. The School of Public Health offers interdisciplinary programs in Health Administration, Maternal and Child Health, Nutritional Sciences, Pathobiology, and Public Health Genetics. More than 50 centers and institutes bring together faculty from throughout the School to collaborate and do research across disciplines.

4.3.5.3.22 Department of Environmental and Occupational Health Sciences

Ranked 5th in the world for environmental and occupational health programs, the Department of Environmental and Occupational Health Sciences (DEOHS) offers graduate and undergraduate degrees in Undergraduate and graduate degrees offered in Environmental Health, Toxicology, Occupational Hygiene, Occupational and Environmental Medicine, Environmental and Occupational Health, Exposure Science and One Health. Department resources includes the Environmental Health Laboratory, the Field Research and Consultation Group, and the Occupational and Environmental Medicine Clinic. Administrative support services are available through DEOHS.

Office: Dr. Kaufman's group has offices available for investigators and study staff in two suites of the University of Washington (UW)'s Roosevelt 1 Building (12 offices and 8 work stations). The Roosevelt building is within a quick shuttle ride of the UW Medical Center and the UW School of Public Health. The departmental office recently moved to enhanced office space in the University of Washington's new Population Health Facility (the Hans Rosling Center for Population Health).

Computer: In the Department of Environmental and Occupational Health Sciences (DEOHS), IT services are provided by a 4.0 FTE team of staff with extensive experience in information technology including, among other areas: systems administration; networking; systems security; database management; systems analysis; application development; and desktop computer support. A set of core services are provided in DEOHS that are available to all projects, and this set includes:

- A centralized account system that supports system authentication and authorization for most systems. Network-attached storage services that provide both individual home directories and access-controlled folders/storage for projects.

- Stringent network firewalls governing access to all Department servers.

- Primary and backup server rooms located in two separate University buildings; server rooms are climate-controlled, access-controlled, monitored, and provided with electrical service protected by uninterruptible power supplies (UPS) and by building generator for some core systems.

- Backup and recovery services that include: "snapshot" document versioning on primary storage system; nightly synchronized backup from primary storage to backup server in secondary server room; quarterly, encrypted, disaster-recovery backups mirrored to location in eastern Washington State.

- Secure remote access to home/project storage using a strict, authentication-based network "whitelist". Automated security and application updates for managed desktop Windows PCs and Apple macOS computers.

- An access-controlled private departmental intranet that provides reference materials for subject areas such as computing.

- Core DEOHS services such as authentication services and server room space are available to and utilized by a number of individual Department programs. As example, the DEOHS IT team now manages a number of project-specific systems providing dedicated computing and analytic power to specialized projects, and the team supports the operation of a variety of statistical software such as R, SPSS, Stata, and SAS along with relational databases (MySQL and PostgreSQL) for researchers.

- Several virtualization host servers running the Microsoft Hyper-V virtualization platform; these hosts are capable of hosting a number of "guest" virtual machines running either Linux and Windows operating systems so as to be able to run tools in both environments on an as-needed basis.

- High capacity primary and backup storage systems running the ZFS file system and with a per-system net storage capacity of over 50 Terabytes. The primary system will export file systems and directories to both Linux (NFS) and Windows (CIFS) client virtual machines.

- A small, high-speed, 10 Gigabit-per-second switch fabric to join computational and storage systems for improved performance.

- A small computing cluster ("Brain") running Rocks 7. The cluster currently consists of 16 compute nodes, and 1 head node with a total capacity of 928 cores, and approximately 6TB of RAM. Cluster nodes are interconnected via 2x 10Gb Ethernet.

4.3.5.3.23 NATIONAL JEWISH HEALTH CAMPUS

National Jewish Health is an academic medical research facility located in Denver, Colorado with nationally recognized research programs in respiratory disease, cardiology, gastroenterology, oncology, allergy, and clinical immunology. Founded in 1899 as a nonprofit hospital, National Jewish Health remains the only facility in the world dedicated exclusively to these disorders. The Institute for Science and Medicine rated National Jewish Health among the "top 10 independent biomedical research institutions of any kind in the world", and the only one that also provides patient care. The mission of National Jewish Health is to make and rapidly translate genomic discoveries into advances for human health so as to bring precision medicine to the forefront of healthcare delivery. National Jewish Health leverages its genomics and proteomics expertise and capabilities to deploy "multi-omic" solutions to accelerate the development of new medical diagnostics and innovative therapies for the prevention and treatment of human disease. At present, there are approximately 230 clinical and basic science researchers spread across three research departments on the main campus of National Jewish Health. Over the past nine years, researchers at National Jewish Health have successfully competed for over $413 million of research funding from over 20 federal funding agencies. Internally, National Jewish Health has several ongoing seminar series to facilitate the exchange of information across multiple research and clinical areas. In addition, National Jewish Health researchers enjoy strong relationships and active collaborations with over 55 academic institutions across the United States and abroad.

4.3.5.3.24 Clinical Research

National Jewish Health has a vast array of clinical research, including: Data Coordinating Centers (e.g., of COPDGene) operated by the faculty and staff of the Division of Biostatistics and Bioinformatics at National Jewish Health (NJH).

National Jewish has a large outpatient clinic focused on respiratory disease as well as one entire floor of the Goodman Building dedicated to a NIH Colorado Clinical and Translational Sciences Institute (CCTSI) which is part of the CTSA network. More details can be found on the website.

Dr. Bowler has research and office space in the Goodman Building. The Research Coordinators will be located adjacent to Dr. Bowler's office. There are sufficient computer resources for the proposed project. Each of the project investigators has password-protected desktop computers (Mac or PCs) with Microsoft Office 2004 and other necessary software. The National Jewish computer network uses Internet Protocol (IP) routers connected to 10/100 Mbps FastEthernet switches that output data into dedicated fiber optic distribution lines tying the National Jewish complex together.

The Division of Biostatistics and Bioinformatics occupies office space in a contiguous 2466 square feet on the second floor of the B'nai B'rith building at NJH. This space includes the Division's own Data Coordinating Center as well as its conference area. The Division maintains secure (restricted card access), independent server room that houses the Division's numerous servers, a dedicated power supply, emergency lighting, and an environmental monitoring and cooling system.

The Division is directed by Matt Strand, PhD, who also serves as the Manager of the Division's own Data Coordinating Center. The Division has two additional PhD-level faculty biostatisticians, an MS-level biostatistician, a PhD-level web developer and database manager, two research project coordinators who double as data quality administrators, seven programmer/database engineers, one database systems engineer, a research data curator, a technical writer, and a senior program administrator.

The Division of Biostatistics and Bioinformatics provides expertise in the design and analysis of clinical trials for investigator-initiated protocols by developing consistent, reliable procedures for the capture, management, and control of study data. Its own Data Coordinating Center, a key component of the Division, operates under Standard Operating Procedures developed in conformance with Food and Drug Administration regulations and Good Clinical Practice guidelines. The Data Coordinating Center collects and manages research and clinical data using Cardiff® TeleForm® technology, as well as its own copyrighted Study Design by Metadata© system. These state-of-the-art approaches accommodate paper, PDF, or Internet-based data collection forms. Data collection forms submitted to the Data Coordinating Center are verified, processed, and stored in a secure database that is accessible to authorized users only.

The Division of Biostatistics and Bioinformatics fulfills other statistical research and educational support functions such as collaboration with clinical and other scientific investigators to design, analyze, and present study results. The Division also assists researchers with protocol development and review through the Clinical Translational Research Center based at the University of Colorado Denver.

All statisticians are equipped with standard statistical analysis programs including the most current versions of SAS, JMP, S+, and R. Flash Drives, CD-ROMS, DVD+RW's and the NJH high-speed network are used for data transfer. In addition, members of the Division are equipped with Microsoft Office Suite, including Access, SQL, BarTender, Adobe Acrobat, Symantec, Quickbooks, Linux, Cardiff® Teleforms, UltraEdit, Thawte and Quality security software, RightFax, Cold Fusion, and ARCserve backup system.

4.3.5.3.25 Computer and network

Dr. Bowler and his team have access to the High-Performance Linux Computing Cluster (S10 RR031832-01 Supercomputer Linux Cluster for Genomics and Proteomics) within the Center for Genes, Environment, and Health at National Jewish Health. The cluster is comprised of a combination of Dell PowerEdge server types and chipsets to address the trade-off between memory and processor intensive jobs, with a total of 432 core processors, a total system memory capacity of 2.4 TB, and secure attached network storage of 200TB with tape-archival capabilities. There is an array of genetic and genomics analysis programs installed on the cluster.

The infrastructure at National Jewish Health is comprised of Cisco and Dell Force10 10 Gigabit Ethernet core switches with all servers being networked with either Gigabit Ethernet or 10 Gigabit Ethernet. There are redundant 10 Gigabit Ethernet connecting IDF closets directly into the core network, providing high speed, low latency data transfers between scientific instruments and the data centers. All client drops are Gigabit Ethernet. National Jewish Health has implemented a Cisco wireless network utilizing 802.11a/b/g/n capable wireless access points, allowing for wireless transfer rates of up to 300mbps. Wireless authentication is performed via Active Directory for internal users, and a wireless network is available for guest users at all National Jewish Health offices and laboratories.

File sharing is accomplished with a variety of servers providing access to data for collaborators and partners through ubiquitous protocols such as FTP, FTPS, HTTP, HTTPS, SFTP, Google Drive, and Dropbox.

National Jewish Health takes a multi-tiered approach to providing proper levels of confidentiality, availability, and integrity for its information technology infrastructure.

For centrally managed host-based anti-malware, intrusion detection, and firewall capabilities, National Jewish Health utilizes Symantec Endpoint Protection on workstations and servers. National Jewish Health utilizes Barracuda Spam Firewalls for anti-spam and anti-malware blocking purposes on the network.

National Jewish Health utilizes Microsoft Forefront Threat Management Gateway (TMG) 2010 for reverse proxying of published web servers. Forefront TMG allows for granular control of published pages on particular servers as well as basic in-line detection and blocking of malicious web traffic. Site-to-Site VPNs are supported. Remote Access (User-based) VPN terminations are currently provided by Cisco ASA 5500 series firewalls, with authentication provided by a Cisco Secure Access Control Service (CS-ACS) server.

Office Space: Over 100,000 square feet of office space is available in the Goodman Smith building at National Jewish Health, including office equipment rooms, conference rooms, offices, cubicles, bioinformatics rooms, and reception areas. Dr.

Bowler has personal office space of approximately 150 square feet located on the seventh floor of the Goodman Building on the National Jewish Health main campus. The office is located within the laboratory. Desk space is available for staff adjacent to their wet bench areas.

Dr. Bowler's primary personal computers are an iMac, a MacBook Pro laptop computer, and a MacBook desktop computer (3.5 GHz 6-Core Intel Xeon E5 with 32 GB 1866 MHz DDR3 ECC memory. All staff have access to other Mac and Windows based computers. All computers in the laboratories are networked through the National Jewish Health high-speed network. Molecular analysis and statistical software licenses resident within the Dr. Bowler laboratory include up to date DNA sequence and protein sequence analysis software, Accelrys MacVector DNA and Protein Sequence Analysis software, Adobe Acrobat Pro, EndNote X7, and others.

Regulatory Compliance: The Office of Academic Affairs at National Jewish Health ensures compliances with federal, state and local regulations with regard to research, creates and supports an environment that promotes safe and responsible conduct in scientific research, and develops and implements effective educational programs that enhance the responsible conduct of research. National Jewish Health scientific staff are trained in human subject research, animal research, recombinant DNA use policies, responsible conduct of research, and laboratory notebook documentation.

Protocol Development: The Director of the Office of Academic Affairs and the Clinical Research Coordinator will provide assistance with the development and submission of protocols, amendments and annual renewals to the National Jewish Health Institutional Review Board (IRB) for the proposed studies under this project.

Biosafety: The Biosafety program at National Jewish Health monitors and enforces Biosafety regulations, directs hazardous waste removal operations with facilities support, set policies in accordance to local, federal and state regulations, oversees all aspects of biohazardous and chemical safety in the laboratories, and provides training to scientific staff in all aspects of laboratory safety.

The combination of state-of-the-art technologies, cutting-edge research programs, strong local, national, and international collaborations, and an unfailing commitment to excellence and urgency in conducting translational research makes National Jewish Health a unique scientific environment in which to conduct the proposed studies. The scientific environment at National Jewish Health promises a very high probability of success in carrying out the proposed work.

4.3.5.4 Data Transfer Capabilities

TB. Fine if to/from BME/Mudd Building but slow from BME/Mudd to CUMC/PH9E015. We use sFTP in general.

4.3.6 Internal & External Funding Sources

- R01-HL077612-15 (2004-27)
- R01-HL093081-12 (2008-24/renewal in preparation)
- R01-HL121270-8 (2012-24/renewal submitted)

4.3.7 Resource Constraints

Slow local transfer speeds. Continuing challenges with cloud-based environment (bureaucratically, data transfer agreements, limited software, limited reliability, difficult of move) in addition to a rental rather than an ownership model of our own data (e.g., we might not be able to afford to use our own data under some models and there is little ability to budget/plan with them).

4.3.8 Ideal Data Architecture

Really fast intranet to pool university computing resources.

4.3.9 Outstanding Issues

Cloud-based solutions continue to have inefficiencies, cost challenges and philosophical issues as mentioned above so there is likely to be a long-term need for a better integrated, faster local systems.

4.4 Dr. Raul Rabadan, Rabadan Lab

Content in this section authored by Dr. Raul Rabadan, Columbia University

4.4.1 Use Case Summary

The Program for Mathematical Genomics (PMG) is a cross campus interdisciplinary effort that brings together evolutionary biologists, computer scientists, physicists and mathematicians, to uncover the structure of genomic data and to study the maps that link genotype to phenotype. It constitutes a space for the free exchange of ideas and methods of quantitative minded scientists with a goal to provide a quantitative understanding of biological systems.

4.4.2 Collaboration Space

We use large amounts of data from our team, collaborators, and public sources.

4.4.3 Instruments & Facilities

CPU and GPU clusters, AWS

4.4.4 Data Narrative

Scientific research

4.4.4.1 Data Volume & Frequency Analysis

Currently the Ramadan Lab produces several PB of data on an annual basis.

4.4.4.2 Data Sensitivity

Some of the data that the Rabadan Lab produces can be considered sensitive, but there are appropriate controls in place for dealing with this data.

4.4.4.3 Future Data Volume & Frequency Analysis

The Rabadan Lab anticipates continually producing PB of data on an annual basis in the future.

4.4.5 Technology Support

4.4.5.1 Software Infrastructure

No answer was provided to this section.

4.4.5.2 Network Infrastructure

No answer was provided to this section.

4.4.5.3 Computation and Storage Infrastructure

No answer was provided to this section.

4.4.5.4 Data Transfer Capabilities

No answer was provided to this section.

4.4.6 Internal & External Funding Sources
No answer was provided to this section.

4.4.7 Resource Constraints
No answer was provided to this section.

4.4.8 Ideal Data Architecture
No answer was provided to this section.

4.4.9 Outstanding Issues
No answer was provided to this section.

4.5 Oncology Precision Therapeutics Imaging Core (OPTIC)

Content in this section authored by Christopher Damoci, Herbert Irving Comprehensive Cancer Center

4.5.1 Use Case Summary

With the heavy emphasis on mouse tumor modeling and experimental therapeutics in the HICCC, OPTIC provides a critical capability to monitor tumor growth longitudinally, and to measure functional aspects of tumor biology. Importantly, all of the imaging instruments in this core facility exist within the animal barrier of the ICRC building. Therefore, mice may be moved freely between holding cages and instruments, a critical capacity for longitudinal imaging studies of tumor growth and progression. This is in contradistinction to other imaging instruments at Columbia University Medical Center which are located outside of the barrier animal facilities. In addition to all of our imaging modalities, the precision therapeutics and personalized models service provide a unique new set of capabilities for HICCC investigators that are closely aligned with the HICCC core focus on translational research and precision medicine. We are the sole centralized resource for in vivo cancer, providing a complete set of preclinical services to facilitate the translation of therapeutic agents and devices to the clinic. OPTIC assists HICCC investigators with preclinical oncology study design and regulatory compliance. Examples of available services include assessment of drug safety and efficacy, discovery of biomarkers of tumor response, preparation and tissue banking of patient derived tumor xenografts, organoids, and cell lines, drug toxicology, PK/PD and antitumor efficacy of single and combination therapies using mouse models of established tumor cell xenografts, patient derived xenografts, and organoids.

The widespread availability of anatomical imaging modalities has enabled longitudinal measurements of tumor growth in preclinical studies, avoiding the need to analyze large cohorts at multiple timepoints and, increasing experimental power while reducing the number of research animals required. Observing the dynamics of tumor growth and treatment response to treatment provides rich information about tumor and drug mechanisms while avoiding the many assumptions required from single timepoint analyses. OPTIC is of use from the vantage of a cancer researcher, grounded in biological questions in which different imaging modalities are employed to yield anatomical, functional, or molecular information about the target. Different applications are optimized to detect and measure the dynamics of tumor initiation, invasion, dissemination, and metastasis in various organs. Primarily, OPTIC provides access and technical support in the operation of these instruments, as well as consultation and training services that add value and accessibility. OPTIC also organizes equipment demonstrations and seminars on both existing and new platforms and services, to inform its users and the HICCC community at large on our capabilities and on emerging imaging technologies for cancer research. Over the current project period, the capabilities of OPTIC were utilized by 43 HICCC members (And a total of 79 Columbia University Medical Center members), supported key data and insights in 25 total peer-reviewed publications all of which were from HICCC Members.

OPTIC supports HICCC research by providing equipment that is up to date, suited to user needs, and well maintained; and by implementing usage policies that ensure that every user is well-trained and supported throughout a project. OPTIC's overall goal is achieved through the following interrelated Specific Aims:

Specific Aim 1: Imaging and analysis. Continue to provide a premier level of imaging and analysis training in a cost-effective and efficient manner to all interested users.

- Maintain, operate and support 8 advanced imaging and blood analysis instruments and a multitude of imaging analysis suites.
- To provide state-of-the-art imaging and analysis with superior convenience, cost-effectiveness and turnaround time for users at any level of experience.

Specific Aim 2: Pre-Clinical Experimental Therapeutics Study Performance. To provide preclinical services to facilitate researchers that do not have the in vivo experience to further their research on their own.

- To design and execute bespoke translational therapeutics studies in a range of murine cancer models including cell line-based xenografts and syngeneic allografts, genetically engineered models (provided by users), and organoid-derived implantation models, and patient-derived xenografts.
- To carry out pharmacokinetic, pharmacodynamic, tolerability, efficacy, and mechanism studies in cancer models and provide users with publication-quality tumor growth data, imaging data, and toxicity data, as well as tissue sample sets for downstream analyses.

Specific Aim 3: Patient-Derived Xenograft Production and Banking.

- Provide a simple process for users to access patient-derived models and samples by enacting a central regulatory process with IRB and IACUC protocols established by OPTIC
- Characterize personalized models with clinical/epidemiological annotation and genomic/transcriptomic analyses, as well as growth kinetics, histopathology, and basic treatment response metrics.

4.5.1.1 Remote/Local Labs

Many labs are located nearby (within a two-block radius) but there are some labs that are downtown and we utilize remote viewing software for a number of them to perform their analysis on our analysis workstations.

4.5.1.2 Data Lifecycle

In regard to all imaging data (and there is a lot of it / large file sizes) Labs are educated that their data is their responsibility to take with them when they are don with their imaging or analysis, usually done through USB hard drives. OPTIC however keeps all data backed up in monthly installments for all of our instruments via backup drives that are exported data from the instrument computers. We have over ten years of data stored this way.

4.5.2 Collaboration Space

Most collaborators/users are within a two-block radius of our site; however, some are at downtown campus or outside entities, but most are within the Manhattan Island.

Data sets are usually the user's responsibility to take with them, but we back everything up every 30 days or so onto external hard drives as a hard backup. In regard to sending data, we occasionally will use our group Dropbox account to send data over to the users and collaborators as requested.

We would really be looking at getting a dedicated internal server for data backup storage and the ability to send files to the users from there. That would be more ideal than the current plan and is a future expansion for us.

4.5.3 Instruments & Facilities

Perkin-Elmer IVIS Spectrum Optical Imaging System (two instruments): The IVIS Spectrum is a state-of-the art instrument for whole animal fluorescence and luminescence imaging. This enables sensitive in vivo detection and quantification of optical signals from engineered reporter alleles engineered into whole animals, into implanted tumor cells, or in microbial species administered to the mice.

Spectral Instruments Imaging AMI HTX X-Ray Optical Imaging System: Engineered with superior optics and industry leading technology, the AMI HTX provides unrivaled sensitivity for bioluminescence, fluorescence and X-ray in vivo imaging. The system a robust, highly effective LED illumination source, customized filter options, and absolute calibration.

Perkin-Elmer Quantum FX micro-CT: This instrument incorporates highly sensitivity X-ray detector in order to enable low-dose longitudinal X-ray computed tomography imaging.

VisualSonics Vevo 2100 High Resolution Ultrasound: Experiments with HICCC investigators have included studies of the pancreas, stomach, esophagus, prostate, bladder, kidney, liver, and eye, as well as longitudinal imaging of transplanted tumors such as xenografts and allografts (both ectopic and orthotopic).

Bruker BioSpec 94/20 9.4 Tesla MRI: This powerful instrument is capable of both high-resolution anatomical imaging and a diverse range of functional imaging applications. This instrument uses advanced functional imaging applications, such as: contrast MRI, diffusion weighted imaging, and spectral imaging.

4.5.4 Data Narrative

With the heavy emphasis on mouse tumor modeling and experimental therapeutics in the HICCC, OPTIC provides a critical capability to monitor tumor growth longitudinally, and to measure functional aspects of tumor biology. Importantly, all of the imaging instruments in this core facility exist within the animal barrier of the ICRC building. Therefore, mice may be moved freely between holding cages and instruments, a critical capacity for longitudinal imaging studies of tumor growth and progression. This is in contradistinction to other imaging instruments at Columbia University Medical Center which are located outside of the barrier animal facilities. In addition to all of our imaging modalities, the precision therapeutics and personalized models service provide a unique new set of capabilities for HICCC investigators that are closely aligned with the HICCC core focus on translational research and precision medicine. We are the sole centralized resource for in vivo cancer, providing a complete set of preclinical services to facilitate the translation of therapeutic agents and devices to the clinic. OPTIC assists HICCC investigators with preclinical oncology study design and regulatory compliance. Examples of available services include assessment of drug safety and efficacy, discovery of biomarkers of tumor response, preparation and tissue banking of patient derived tumor xenografts, organoids, and cell lines, drug toxicology, PK/PD and antitumor efficacy of single and combination therapies using mouse models of established tumor cell xenografts, patient derived xenografts, and organoids.

4.5.4.1 Data Volume & Frequency Analysis

Currently OPTIC produces GB scales of on an hourly basis.

4.5.4.2 Data Sensitivity

Some of the data that OPTIC produces can be considered sensitive, but there are appropriate controls in place for dealing with this data.

4.5.4.3 Future Data Volume & Frequency Analysis

OPTIC anticipates producing GB scales of data on an hourly basis in the future.

4.5.5 Technology Support

4.5.5.1 Software Infrastructure

Dropbox is the primary software used for the business as well as TeamViewer and zoom

4.5.5.2 Network Infrastructure

We work through our IT and a number of our instruments are connected to the internet for mostly ability to check email or transfer some files to Dropbox, but that is all currently. We would like to better utilize these connection opportunities in some fashion.

4.5.5.3 Computation and Storage Infrastructure

There is no on-site computation and storage within OPTIC.

4.5.5.4 Data Transfer Capabilities

Sizes are normally int the range of 1-10GB of data, We utilize Dropbox for the share of the data.

4.5.6 Internal & External Funding Sources

We are supported by the HICCC cancer support grant, but the majority of our funding comes from the user utilization of our instruments or processes.

4.5.7 Resource Constraints

Our biggest concern is the safety of our hard backups that are aging and having a storage solution on the cloud for them.

4.5.8 Ideal Data Architecture

Having a server that would be a secondary backed up storage with also a cloud backup (like Backblaze, or similar).

4.1.9 Outstanding Issues

No additional issues were reported.

4.6 NewYork-Presbyterian Hospital Technology Support

Content in this section authored by Zhani Pellumbi and Charles Corona, Columbia University Irving Medical Center and Mark Turczan NewYork-Presbyterian Hospital

4.6.1 Use Case Summary

This case study provides the state of the NYP network in its entirety as it currently stands (June, 2024). To the best of our knowledge any current research computing takes place over the "commodity" network, designed to support the typical application communication needs across the entire NYP Enterprise.

4.6.2 Collaboration Space

The IT organizations collaborate to support the research, clinical, and enterprise use cases of the campus.

4.6.3 Capabilities & Special Facilities

Core Resources provides all network connectivity between each campus building, between campuses, to datacenters and to the Internet. All of the inter-building/campus/datacenter links traverse our internal maintained fiber infrastructure(both multimode and singlemode), and our provider maintained dark fiber pathways. Core Resources maintains the interconnects where the dark fiber terminates. In addition to network connectivity, Core Resources also maintains and manages network services such as NTP\DDI for the whole of the enterprise.

4.6.4 Technology Narrative

4.6.4.1 Network Infrastructure

The campus LAN follows a tiered architecture. Within a building, each floor(Layer 2) collapses to redundant building distributions(Layer 3) which in turn connect to redundant cores at both of our main campuses in New York City. Each campus connects to each other, our datacenters, and our regional hospital network over our dark fiber infrastructure. ""Commodity internal network connectivity"" (think mouse flows) is provided with 1Gbps connections to devices and (for the most part) 10Gbps uplinks to the building distributions. Between the building distributions, cores, and across the WAN with redundant 10Gbps connections. Smaller doctors' offices/practices are connected via EVPL or via secure tunnels through the Internet.

The NYP datacenters follow a CLOS architecture. They are interconnected with a redundant 100Gbps connection. NYP reaches the Internet via commodity carriers and tier 1 providers along with Research Internet via Internet 2 services. NYP leverages Express Route to connect to Microsoft Azure cloud capabilities. CUIMC leverages AWS Cloud capabilities 10Gbps Megaport SDN connections. There are Globus services in AWS as well.

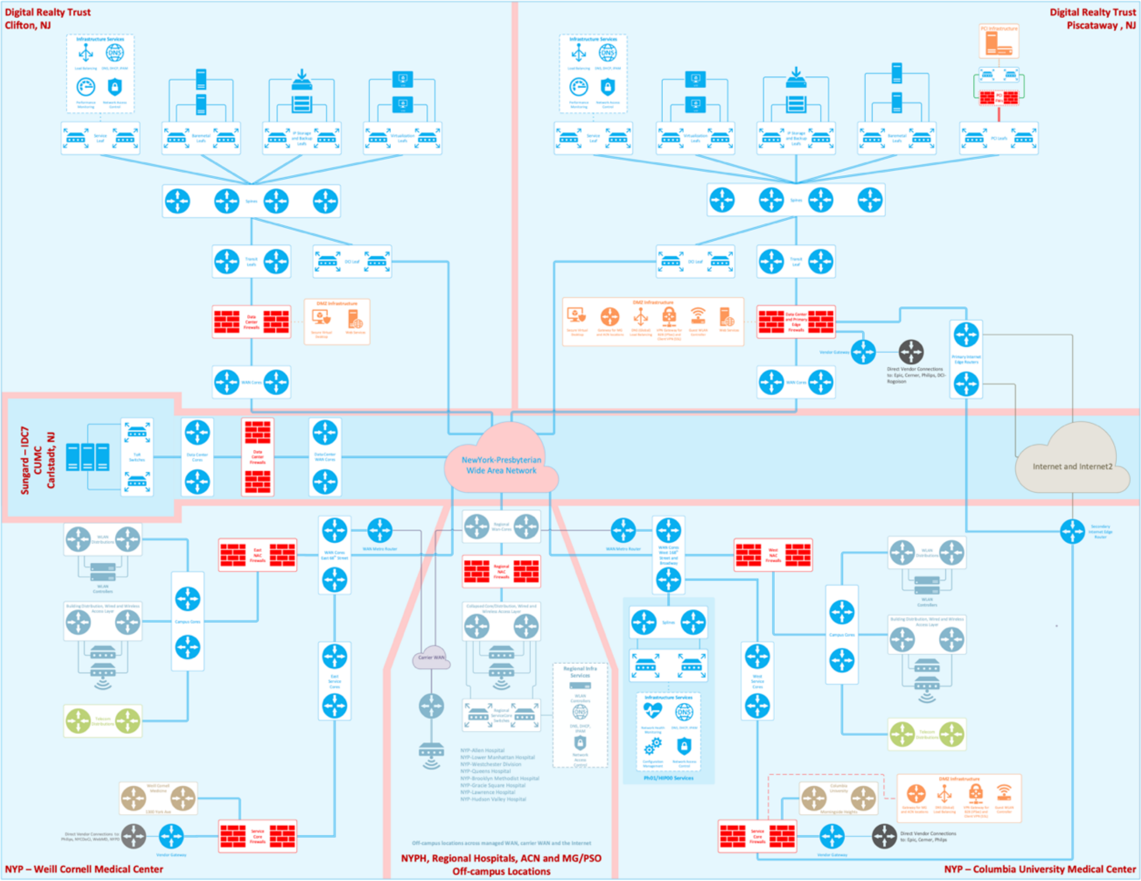


Figure 4.6.1 – NYP Network Diagram

4.6.4.2 Computation and Storage Infrastructure

We have a large storage environment for scanned digital pathology image slides. We have deployed 2PB year one and will grow at about 2PB per year. This will also have needs for image scanners to be able to send data from the campus network where they are deployed at 1GPBS/scanner. Phase II will be viewing software, Imange Management System, which will be cloud based, PathAI.

We have deployed 2PB of Archive Flash storage, based on Pure Flashblade environment. This is deployed in our Datacenter in DRT and accessed from VM servers in DRT over SMB Protocol. Also, on the campus we are deploying Aperio GT450 Scanners, 5 so far 15 more to go. Each scanner bursts 1GPBS and streams scanned images to the datacenter storage. We are also replicating all this data to the secondary data center in DEL, where we have also deployed 2PB of Pure FlashBlade.

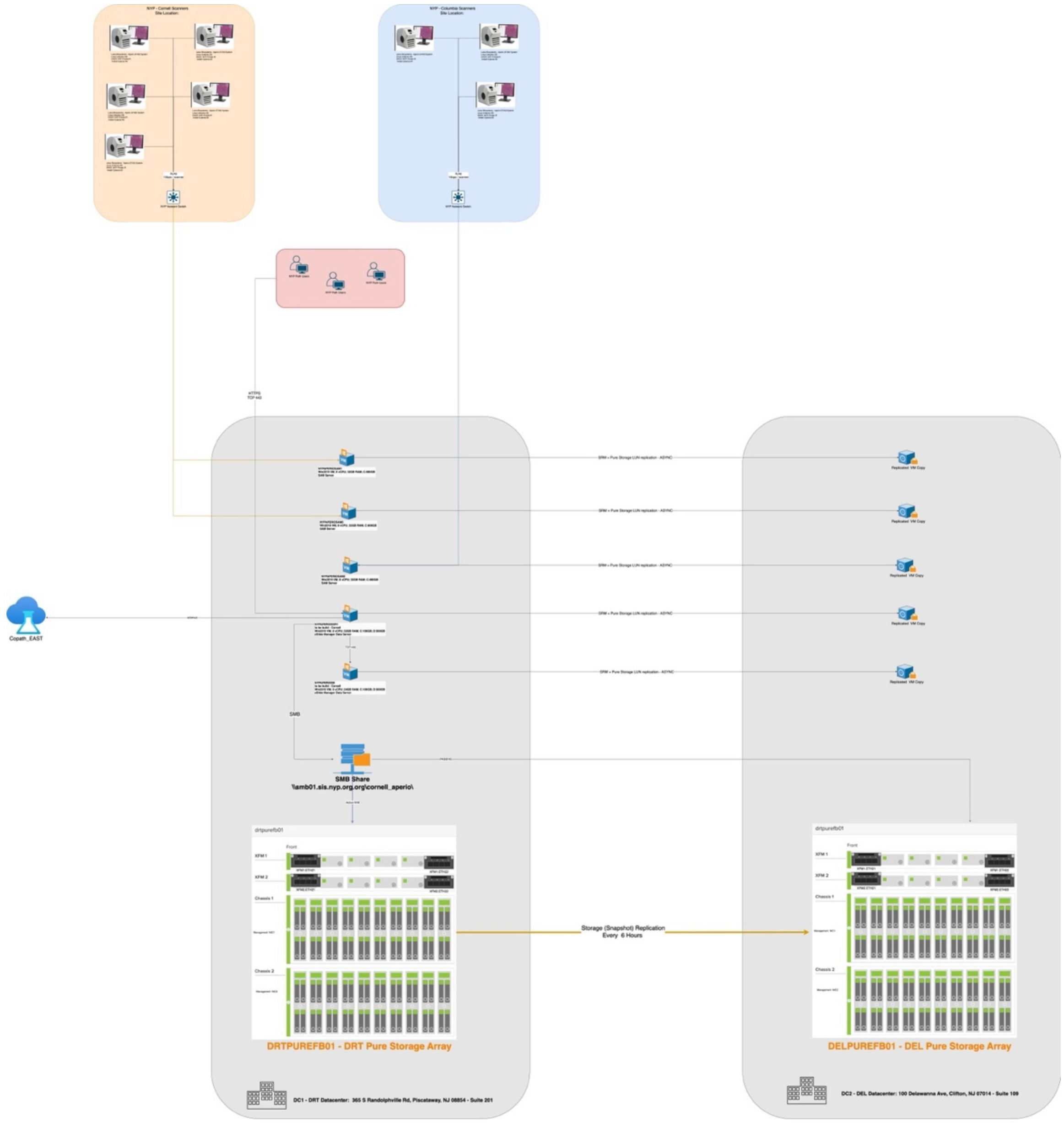

NYP - Cornell Scanners
Site Location:
NYP Network Switch
NYP - Columbia Scanners
Site Location:
NYP Network Switch
NYP Path Users
NYP Path Users
NYP Path Users
HTTPS
TCP 443
SRM + Pure Storage LUN replication - ASYNC
SRM + Pure Storage LUN replication - ASYNC
SRM + Pure Storage LUN replication - ASYNC
SRM + Pure Storage LUN replication - ASYNC
SRM + Pure Storage LUN replication - ASYNC
Replicated VM Copy
Replicated VM Copy
Replicated VM Copy
Replicated VM Copy
Replicated VM Copy
Copath_EAST
SMB
SMB Share
\\smb01.sis.nyp.org.org\cornell_aperio\
drtpurefb01
Front
XFM 1
XFM 2
Chassis 1
Chassis 2
Storage (Snapshot) Replication
Every 6 Hours
drtpurefb01
Front
XFM 1
XFM 2
Chassis 1
Chassis 2
DRTPUREFB01 - DRT Pure Storage Array
DELPUREFB01 - DEL Pure Storage Array
DC1 - DRT Datacenter: 365 S Randolphville Rd, Piscataway, NJ 08854 - Suite 201
DC2 - DEL Datacenter: 100 Delawanna Ave, Clifton, NJ 07014 - Suite 109

Figure 4.6.2 – Leica Biosystems Aperio GT450 at Weill Cornell

4.6.4.3 Network & Information Security

There are firewalls for the entire Enterprise network, datacenters, and campus locations. The campus network also leverages both wired and wireless endpoint NAC solutions. We use the Medigate platform to identify end devices. All wired end devices need to be registered within IPAM to obtain IP addresses, and the wireless side of the house uses certificates to authenticate valid devices.

4.6.4.4 Monitoring Infrastructure

LogicMonitor is the current monitoring platform for the Enterprise. It is attempting to replace CA Spectrum as that product is being sunset. Arista CVP does gather device statistics and network flows. We currently leverage netMRI for cross-vendor device configuration management. We use Splunk dashboards to provide insight into syslog messages generated by Cisco network devices as well. We will be deploying perfSONAR for the Science DMZ and Enterprise.

4.6.4.5 Software Infrastructure

Software is maintained by individual researchers and labs.

4.6.5 Organizational Structures & Engagement Strategies

4.6.5.1 Organizational Structure

The NYP Core Resources team manages all of the network infrastructure(LAN/WAN), inclusive of firewall, load-balancing, DNS, and DHCP services for both the NewYork-Presbyterian Hospital Enterprise(NYP) and Columbia University Irving Medical Center(CUIMC). We also interconnect with WeillCornell Medicine. Their teams manage their own network/compute/storage infrastructure.

4.6.5.2 Engagement Strategies

There are infrastructure teams in both organizations(NYP/CUIMC) who manage their own compute and storage services, both centralized and departmental IT teams in CUIMC who manage individual servers/workstations

4.6.6 Internal & External Funding Sources

Nothing to report in this section, information technology is centrally funded.

4.6.7 Resource Constraints

Nothing to report.

4.6.8 Outstanding Issues

Nothing to report.

4.7 Columbia University Irving Medical Center Cloud Services

Content in this section authored by Zhani Pellumbi and Charles Corona, Columbia University Irving Medical Center and Mark Turczan NewYork-Presbyterian Hospital

4.6.1 Use Case Summary

The following profile will describe some of the cloud services available to the CUIMC community. CUIMC-IT provides a HIPAA compliant AWS infrastructure and soon (available approximately 2024/6) a HIPAA compliant Google GCP infrastructure. The CUIMC GCP right now carries non-HIPAA workloads and Terra.BIO workloads.

4.6.2 Collaboration Space

The services provided are available to all researchers.

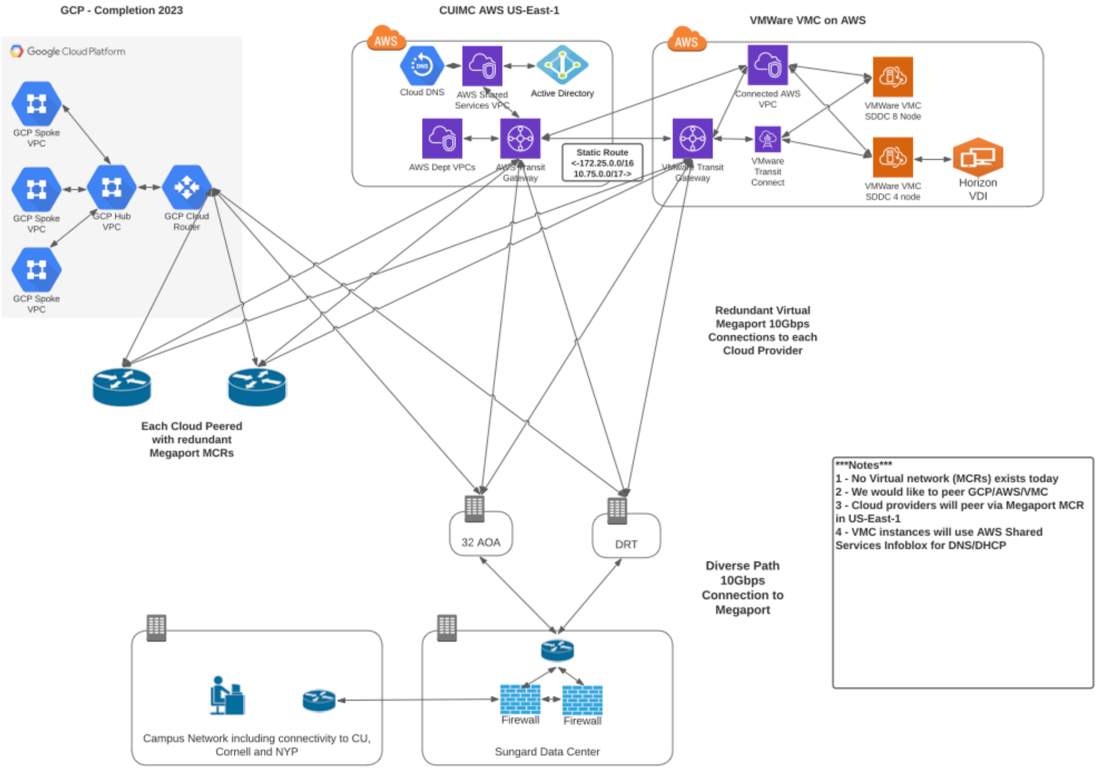
4.6.3 Capabilities & Special Facilities

AWS, GCP and VMware VMC on AWS infrastructure. Every CUIMC AWS receives significant discounts from the NIH sponsored STRIDES program. On GCP, only researchers with STRIDES grants receive these discounts.

4.6.4 Technology Narrative

4.6.4.1 Network Infrastructure

CUIMC has two main campus network connectivity points to CUIMC AWS and GCP. The first and main connections are multiple 10Gbps Megaport connections that are split between AWS and GCP. The second is connectivity via Internet2.



Notes
1 - No Virtual network (MCRs) exists today
2 - We would like to peer GCP/AWS/VMC
3 - Cloud providers will peer via Megaport MCR in US-East-1
4 - VMC instances will use AWS Shared Services Infoblox for DNS/DHCP

Figure 4.7.1 – NYP Access to Cloud Resources

4.6.4.2 Computation and Storage Infrastructure

There are a range of different compute and storage services that are being used on CUIMC AWS and GCP by a number of members of the CUIMC community.

On AWS, S3 and Glacier are commonly used for mass object storage for not only analysis but for archival storage. S3 is often used to facilitate data transfers between CUIMC and external organizations, once the transfer is approved. Also in use are AWS based NetApps, Nasuni Network filers, AWS File Gateways and AWS Windows Fsx file servers. These storage solutions are accessed from both campus and AWS. The data is generated on campus, in the cloud and off-site. We are utilizing Globus for intra and inter-organizational data transfers.

Compute use on AWS is based on several different technologies. EC2 instances such as on-demand and spot instances are used for a variety of workloads including workloads that do analysis using Python, Matlab or R. Databases services on EC2 instances or using AWS RDS such as MySQL, PostgreSQL and SQL Server are commonplace. Container workloads are using AWS EKS (Kubernetes), Docker and AWS ECS. Many common use cases such as AWS Parallel Cluster (HPC) and Memverge use ephemeral/spot instances that once processing has completed, they are deleted. These instances sometimes will use GPUs such as the NVIDIA A100/H100 which are spun up and shutdown. Also in use are FPGAs that are utilized by applications such as Illumina's Dragen software which is used for analysis of sequencing data. A Ronin HPC environment also runs on the infrastructure (https://ronin.cloud/). AWS Sagemaker machine learning environment is also used. Serverless Lambda functions are also being used. AWS HealthOmics is being used to analyze genetic data. RedCAP and AWS SFTP are other common applications.

In addition to those workloads above, Citrix Virtual Desktop solutions and a 3CX PBX have been deployed.

Running on top of the CUIMC AWS infrastructure is VMware VMC (https://aws.amazon.com/vmware/). All VMware virtual machine instances that ran in the CUIMC SunGard/365 datacenter on the Dell VxRail VMware infrastructure have been migrated to the AWS VMC infrastructure. In addition, a Horizon VDI solution has been deployed to allow for a secure virtual desktop environment.

GCP is in limited use at CUIMC until the full deployment of the HIPAA Secure Data Enclave is completed in 2024/6. Currently GCP is being used for mass object storage using Google Cloud Storage and more importantly for using the NIH's variety of tools that can be accessed via their AnVIL website such as the Broad's Terra.Bio is used for data analysis and SEQR (https://anvilproject.org/)

4.6.4.3 Network & Information Security

Connectivity to the CUIMC cloud infrastructure (AWS/GCP/VMC) is via multiple 10Gbps connections. One is an AWS Direct Connect and the other two are provided by Megaport. Megaport allows us to virtualize our connectivity to the cloud providers. In addition, We will use Megaport to peer AWS and GCP in their local locations in Virginia/South Carolina which will save on bandwidth since this traffic will not need to traverse back to campus. AWS VPCs are using the AWS Transit Gateway that is set up in a hub and spoke model. GCP uses a VPC/subnet and a GCP Cloud Router.

4.6.4.4 Monitoring Infrastructure

CUIMC uses AWS Cloudwatch/Cloudtrail and GCP Cloud Monitoring for systems monitoring. VPC Flowlogs are used for network analysis. For AWS security monitoring, CUIMC uses AWS Cloudwatch/Cloudtrail, AWS Guardduty/SecurityHub, AWS Detective and AWS Inspector. For GCP, CUIMC uses Cloud Monitoring, VPC Flowlogs for network analysis and GCP Security Command Center for security monitoring and alerting.

4.6.4.5 Software Infrastructure

There are many software packages and services in use. Some of them are: AWS Omics, AWS Sagemaker, MySQL/PostgreSQL/SQL Server, Python/Conda/Cuda/PyTorch/Tensorflow, R/R-Studio, AWS Parallel cluster/Ronin/MemVerge, Illumina Dragen, MatLab, RedCAP and AWS SFTP.

4.6.5 Organizational Structures & Engagement Strategies

The AWS Third Thursday series done once a month. Weekly Cloud Users Meeting. One-off meetings with the research community and with cloud providers to help architect solutions or solve issues.

4.6.6 Internal & External Funding Sources

NIH STRIDES

4.6.7 Resource Constraints

The single biggest challenge is CUIMC having inconsistent network throughput. Researchers are often unable to copy/move large chunks of data while generating or analyzing and the data. This results in many hours of lost productivity. In addition, CUIMC is not able to fully utilize our Internet2 connectivity.

4.6.8 Outstanding Issues

Nothing reported in this section.

5 Case Study Discussion & Campus Planning

On June 26th and 27th 2024, staff from NYP, CUIMC, NYSERNet, and EPOC participated in a discussion on the use cases and potential next steps to develop a set of sustainable approaches to provide technological support. Notes from this set of discussions appear in the following sections.

5.1 Dr. Despina Kontos, Computational Biomarker Imaging Group (CBIG)

The Computational Biomarker Imaging Group (CBIG) works with radiologic imaging, clinical epidemiology, biostatistics, and cancer biology ,and how these can be joined with technology advancements in image analysis, machine learning, and artificial intelligence. The overall research vision is to act as a translational catalyst between imaging data science and clinical cancer research.

Dr. Kontos's background is more on the computer science spectrum, and she brings experience in computing imaging, artificial intelligence, and data statistics to many of the projects. The lab deals with a variety of multi-modal data inputs; not all of which can be direct related to observational health data gathered in a clinical setting. This being said, the clinical data that is used comes from others primarily. This lab doesn't design trials or do collection; their work relies on leveraging medical images gathered through other means is more common, and then applying novel detection algorithms (mainly to detect and categorize cancer). The data sets they are dealing with can approach large order TB, especially when they are actively running analysis. This implies the need to store PB of research data into the coming years. Data sets do not lose value over time, and will continue to retain usefulness.

A typical workflow may proceed as follows:

- Identification of data streams (could involve thousands or images)
- Curation of data sets; typically bringing these into the location where computational resources can access them and applying any classifications to the data that are needed. The lab relies on the data being correctly classified and protected:
    - All clinical data has been de-identified by trained staff.
    - All necessary IRB approvals are obtained.
    - Data is managed and stored following the necessary precautions.
    - Trained staff are available to follow-up throughout the research process.
- Perform analysis using computational resources (local, cloud, wherever they can be found)
- Build models based on the initial analysis, work with collaborators to verify the models.
- Make recommendations back to the clinical side of the research ecosystem.

Computational resources are a challenge to maintain due to requirements and cost. Dr. Kontos used resources at previous institution (University of Pennsylvania) that were located on-site for a monthly fee. The facilitates were HIPPA compliant, and included storage, commutation, and network bandwidth. Currently the lab uses resources at the Center for Computational Biology and Bioinformatics (C2B2)[3]. One area of friction is the limited bandwidth, and the length of time needed to transfer research data. Due to these challenges, they have explored the used of cloud resources for data storage and transfer but found those to be no better.

The lab would find it desirable to see the following changes implemented:

- A model co-op computation and storage would work better for their own use case: they would pay for resources on a regular basis, if they could occasionally "burst" above their share for particularly large jobs (and allow their paid for resources go towards other use cases when they are not using them).
- The ability to have a seamless backup system would also be desirable, as they have no formalized way to manage this now.
- Uniform availability of computation, storage, and networking instead of systems that differ based on location and age. E.g., one method to use computation with a similar architecture and software stack. Along with one method to manage data curation and mobility.

Discussion with the lab produced the following statements of fact, and questions for the future design of the technology support structure:

- Cloud resources are desirable for researchers in the lab, namely because they scale to what is needed in terms of an analysis workflow, the hardware is performant and modern, and the software stack is maintained and easy to use. This is starting to become more desirable than on-site solutions (namely C2B2) because it's almost impossible to "keep up" with the changes in technology and be cost effective. There is a fundamental friction with the use of off-site cloud resources for medical work though: historically the security concerns have favored "keeping the data close", which would imply the hospital environment is better served to maintain its own hardware and software and absorb the costs to support the researchers.
- Most, if not all, of the data that they work with is automatically curated and cataloged: especially for AI and ML use cases. Tools like flywheel[4] are used heavily. One part of the process that is still manual involves the de-identification of data. This is done internally to the hospital and can be slow. Ways to improve this (for everyone in this environment would be beneficial to explore).
- An ideal data architecture (assuming no limitations on cost) should involve:
    - Predictable and dependable ingest procedure for data that would capture the steps of de-identification and classification of data,

[3] https://systemsbiology.columbia.edu/center-for-computational-biology-and-bioinformatics-c2b2

[4] https://flywheel.io/

    methods to appropriately store and backup data sets (as well as applying permissions for who can access)
  - Ways to efficiently retrieve data for use in analysis pipelines.

- Data sharing is currently inefficient and relies on a heterogeneous environment that fits individual researcher approaches. The general lack of high-speed networking is a factor; however, the use of different and ineffective tools makes the issue much more challenging. Web and cloud-based sharing methods (e.g., Dropbox, etc.) are widely used, but break down as the file sizes and data sets grow. Shipping media is still widely used. Relying on "neutral" sharing locations (e.g., AWS) work well when commutation is involved, but do not scale as just a sharing mechanism. For the later to be a solution, entire platforms must be built to offer storage, computation, and access control: this can be expensive and may be out of reach for some research efforts.
- The medical research field is seeing disruption from industry. In particular the rise of basic medical care being offered from pharmacy retail locations, and competition for workforce are making it hard to train and retain staff. This new reality will influence medical research in the coming years.

The discussion ended with thoughts on what a future computational and storage environment (on premises or off) would look like for the research community at NYP and CUIMC:

- Predictability and reliability must be a top priority. This means a uniform environment with expectations for access, speed, efficiency, and support.
- Agnostic on where this is location: cloud or local is not of a concern to some researchers, provided that the data mobility aspects can be transparent. Non-medical resources have taken a hybrid model over the years to integrate the local storage (e.g., golden, and secure data copy) with the resources of clouds and have been successful[5] [6].
- Cloud platforms have an added benefit of making it easier to locate or share research data with other groups that may want to perform similar research.
- No expectation that the services will be free, but straightforward and competitive costs to use.
- Scalable toward the needs of large and small research projects
- Ability to leverage knowledgeable staff resources to adapt workflows.
- Efficient ways to manage data sharing and mobility between resources and collaborators.

[5] https://data.lsst.cloud/

[6] https://next-gen.materialsproject.org/

5.2 Dr. Michal Levo, Gene Regulation and Genome Organization

The Levo lab combines several experimental modalities (from genomic assays to live sub-cellular imaging), with computational analysis and modeling, to study the dynamic, multilayered control of gene expression in the context of genome organization, particularly in the service of differentiation and development. Studies focus on perturbation-based approaches and quantitative measurements to facilitate causal insights and inform predictive models.

Dr. Levo has a background in both biology and computational sciences and hopes to incorporate technology into her research process. As a new hire, it has been challenging to acquire technology support thus far. One area of friction has been the inability to migrate a data set from her prior institution (Princeton University) to NYP/CUIMC resources. A prior attempt to leverage a cloud storage location temporarily (e.g., AWS) was not successful.

Ideally, the Levo lab needs predictable access to computation and storage resources: the location of these is not a major concern (e.g., located in a cloud or on site). The primary concern is availability and capability: once they create analysis pipelines for their work, they want to rely on the technology to be available and efficient to the tasks they have designed. In general, they do not want to build and operate their own computational system (more time working on technology takes away from the ability to do research). In particular, "interactive computing" (e.g., the use of notebook software) could be a valuable way to approach future analysis tasks.

An example of a workflow they will need to support is as follows:

- Use of a clinical instrument (e.g., microscope)
- Creating a pipeline to take the observational output of the instrument and perform local calibration and analysis: this is typically a local PC that is directly connected to the acquisition PC of the instrument.
- TB of information will flow between these machines (e.g., 10+TB of raw images, and 100s of GB of processed images)
- Initial processing can be done on a single machine, ensemble studies of multiple images must be done on larger resources (e.g., C2B2, or cloud resources)

Data mobility will be an important part of the workflow for the Levo lab: transferring data to collaborators or to other locations where computation can be done. Tools like Globus can help but must be available and administered uniformly. Want to prevent a situation where data is "marooned"; see the prior discussion of data that was attempted to be moved from Princeton University. The Levo lab expects to collaborate with entities outside of the NYP and CUIMC ecosystem, in particular there are emerging use cases with researchers at UCSD and others. Establishing known methods to share data are a requirement.

5.3 Dr. R Graham Barr, Respiratory Epidemiology

The Barr Lab is studying the epidemiology of lung structure using large-scale analysis of imaging data of the lungs, mostly using research computed tomography (CT) scans but also cardiopulmonary magnetic resonance imaging (MRI).

Dr. Barr was not available for an in-person discussion, but the following observations were made on the use case:

The sizes of medical imaging data (CT or MRI) vary widely and will increase in the coming years. For workflows that span facilities, a workflow that can accommodate the seamless data transfer between observational machine, local storage, computation, and remote collaboration will be needed. This should include tools that are easy to manage data transfer in a secure and efficient manner.

The collaboration locations are not well known, but once they are, it is in the best interest of NYP and CUIMC to establish communication with the far-ends to test data mobility capabilities between facilities. In some rare instances, a well-known collaboration may benefit from specialized network overlay between facilities[7].

As in the Kontos Lab use case, data de-identification procedures are critical, and may not be efficient when coupled to an automated workflow. Ways to improve this for all clinical data users should be investigated.

[7] https://lhcone.web.cern.ch/

5.4 Dr. Raul Rabadan, Rabadan Lab

The Program for Mathematical Genomics (PMG) is a cross campus interdisciplinary effort that brings together evolutionary biologists, computer scientists, physicists and mathematicians, to uncover the structure of genomic data and to study the maps that link genotype to phenotype.

The core use case for the Rabadan Lab is mathematical genomics, systems biology, and computer science. Recent work has focused on the use of AI/ML (developing LLMs) that focus on castrogenomics, and microbiomes. The lab is attempting to categorize and learn from the outcomes of different treatments on various forms of clinical data and imaging. The data sets they are dealing with can be large, and will grow over time. Early estimates are that the entirety of the catalog and derived products can approach PB scale. The majority of the computational work is built into the cloud (e.g., AWS), although there are local GPU and storage resources that can be leveraged for development and smaller scale work.

Scalability of the available technical resources will be a future factor to consider. Based on current trends, the Rabadan Lab could easily use more GPU and CPU resources in their work if they were available. As a result of this, they typically scale their analysis to use what they know is available, and that can limit efficiency.

The Rabadan Lab does not have a preference for where resources are physically located: the use of clouds or local is not material for the work they are doing. They can be sensitive to cost however and acknowledge that the use of cloud resources can be expensive (both in terms of storage and analysis). This will only increase as the data volumes grow beyond PB, and the time needed to perform analysis with available resources increases.

The Rabadan Lab struggles with acquiring, training, and retaining staff. In particular they want to get new staff members with experience in computational fields and want to show them they have access to leading edge resources. Once they are trained, the lure of industry and higher paying roles can be strong.

The Rabadan Lab uses clinical data but adheres to all procedures for de-identification. In the general case managing this data is not a challenge. In general, they want to consume electronic data that is already curated and de-identified: they are not equipped to create and manage their own sets of data. The individual size of data sets varies highly, and they do leverage community resources where they can[8] [9]. TB scale is common, and these are growing as machines become more accurate. Due to the size (and availability and cost of storage), they are frugal in what is downloaded and how it is used over time.

[8] https://www.ukbiobank.ac.uk

[9] https://jgi.doe.gov/

Their approach to technology has adapted over time, and they value the ability to keep a workflow they have established running efficiently, wherever the resources may be located.

5.5 Oncology Precision Therapeutics Imaging Core (OPTIC)

The Oncology Precision Therapeutics and Imaging Core (OPTIC) was created to monitor tumor growth longitudinally, and to measure functional aspects of tumor biology. All of the imaging instruments in this core facility exist within the animal barrier of the ICRC building. OPTIC is the sole centralized resource for in vivo cancer, providing a complete set of preclinical services to facilitate the translation of therapeutic agents and devices to the clinic. They assist HICCC investigators with preclinical oncology study design and regulatory compliance:

- ***Specific Aim 1***: Imaging and analysis. Continue to provide a premier level of imaging and analysis training in a cost-effective and efficient manner to all interested users.
- ***Specific Aim 2***: Pre-Clinical Experimental Therapeutics Study Performance. To provide preclinical services to facilitate researchers that do not have the in vivo experience to further their research on their own.
- ***Specific Aim 3***: Patient-Derived Xenograft Production and Banking.

OPTIC initially operated small instruments (e.g., MRI, Ultrasound, CAT scan), and dealt with data volumes that could easily be brought out on removable media: < 1GB files, and GB scales data sets. As the use cases and instruments grew, the ability to handle data became more critical. OPTIC can perform in-office analysis by sending observational data from local instruments to resources that are located on-site.

Future requirements for OPTIC include:

- A more efficient way to manage data backups. For now, the use of portable media is common. A tool that can manage this (e.g., Backblaze[10]) would be more efficient in the long run. TB scale is sufficient for now, data may not grow to PB scale in the near or medium term.
- Ability to handle analysis, particularly when the user is offsite (extremely common). Many of the instruments use a proprietary format for the data and have analysis software that requires local access. A system of remote access relying on NoMachine[11] and TeamViewer[12] is available, but not ideal.
- Uniform operating environments. Right now, the systems they support are a mixture of operating systems and ages, and as a result of this security for online systems can be challenging to manage.
- Data growth will only increase, exacerbating the need for storage and networking.

Given that OPTIC deals with non-human data, they are not subject to some of the other requirements on information security as with other NYP and CUICM use cases. This can simplify some of the data handling infrastructure for storage and sharing.

[10] https://www.backblaze.com/
[11] https://www.nomachine.com/
[12] https://www.teamviewer.com/en-us/

5.6 NewYork-Presbyterian and Columbia University Irving Medical Center Technology Support

The technology support structures of NYP and CUIMC (networking, local storage and computation, cloud use cases, security) were combined into a single session of discussion despite the multiple case studies. There were multiple areas of discussion on the case studies themselves, as well as the overall ways the institutional systems can be enhanced to support the research use cases.

NYP and CUIMC presented information on a Digital Pathology Storage Platform, which was designed to simply the workflow surrounding the handling of observational data from tissue samples and digitizing them for use in other parts of the clinical process. The workflow for this is as follows:

- Glass slides are digitized using on-site technology. This results in images that are between 1-50GB each, and they can be processed in batches of between 50 and 100 slides.
- Once digitized, the data lands on local NAS devices (e.g., all scanners are able to communicate with this SAN to simplify the ingest of data).
- Data sets can then be fed into an AI/ML categorization system (PathAI[13]) that is being housed in the cloud (e.g., AWS).
- All analysis of the data is done within the cloud environment.
- Data is returned back to NYP and CUICM resources and can only stay within the cloud environment for approximately 30 days. This results in a local data set that is around 2PB (and growing) and comprised of over 700,000 slides.

This pathology use case (and others like it) rely on a hybrid approach of local and remote resources to be effective. For now, it functions almost as two different workflows (local capture, storage, and then batch transfer and processing). Future use cases may be defined by more on-line analysis that involves a streaming workflow.

The structure of the NYP and CUIMC networks was designed with smaller network flows in mind, and until recently has not considered the establishment of a dedicated Science DMZ. In particular, the security concerns surrounding the clinical and research use cases remains hard to disambiguate, implying that all network traffic must be treated the same way.

Overall, the campus does have adequate connectivity to support a number of use cases (multiple commodity network connections, as well as access to Internet2), but the effective throughput of any one connection will be limited by the capabilities of the internal network. Cloud connectivity is supported by Megaport[14], and is seeing increased use. Internet2 does offer Megaport connectivity, but since both institutions are not direct connectors/members, this may not be an option.

[13] https://www.pathai.com

[14] https://www.megaport.com/

The one area where the NYP and CUIMC networks must grow in the coming years is supporting the various cloud computing use cases that were highlighted in the case studies. Currently, there numerous efforts underway to create HIPAA compliant infrastructures in AWS and GCP (expected by 2026). The campuses offer a VPC for connectivity to the cloud, which has helped to simplify the integration of local and remote resources. Overall, they are seeing the following basic approaches to the use of cloud:

- Backups
- AI/ML experimentation
- Creating micro-services
- Analytics
- File sharing.

For research projects, NYP and CUIMC have noticed that research groups are leveraging a number of the available cloud resources:

- On AWS, S3 and Glacier are commonly used for mass object storage for not only analysis but for archival storage.
- S3 is often used to facilitate data transfers between CUIMC and external organizations,
- Compute use on AWS is based on several different technologies.
- EC2 instances such as on-demand and spot instances are used for a variety of workloads including workloads that do analysis using Python, Matlab or R.
- Databases services on EC2 instances or using AWS RDS such as MySQL, PostgreSQL and SQL Server are commonplace.
- Container workloads are using AWS EKS (Kubernetes), Docker and AWS ECS.

Some of this work is being supported by NIH funding[15], and will result in a more inform way to leverage cloud and local resources for researcher workflows. Running on top of the CUIMC AWS infrastructure is VMware VMC[16]. The data is generated on campus, in the cloud and off-site. We are utilizing Globus for intra and inter-organizational data transfers.

Some of the challenges that NYP and CUIMC have run into since migrating many use cases to the cloud include:

- Inconsistent network throughput remains high. Researchers are often unable to copy/move large chunks of data while generating or analyzing and the data. This results in many hours of lost productivity.
- CUIMC is not able to fully utilize Internet2 connectivity, since they are not directly connected.
- Researchers do not always utilize the cloud resources correctly or consistently. There is a way to use "institutional sponsored" cloud resources, and then there is also going and doing it yourself (which is almost always

[15] https://datascience.nih.gov/strides

[16] https://aws.amazon.com/vmware/

more expensive). For former is certainly preferred. When local staff do catch use outside of their control, they are often able to intervene and mitigate before charges become too great.

5.7 Other Discussion Topics

Computation and storage for researchers, and the best ways to support for near- and medium-term needs, has been identified in all case studies. The campus has a mixture of local and cloud-based options depending on the requirements of the research and availability of funding. A part of the engagement was devoted to ways that local resources could be augmented to support computing needs in the future.

Condo computing[17] is one mechanism that may allow the campus to scale computing needs against what funding is available from individual research groups. For example, a research group may have a certain amount of dollars to spend on CAPEX for computing and storage, that they could use to purchase campus resources. The campus could use this funding to refresh existing hardware (as needed), as well as manage the OPEX (e.g., software, maintenance, power, etc.). Resources are allocated based on purchase power, but can also go above when available. The campuses could offer this model, as well as unifying the way that resources are managed at each location, to unify the available computation and storage ecosystem.

A significant amount of time was also devoted to identifying methods that can be used to implement a "Clinical DMZ" environment to improve data transfers within and external to the campus. The following diagram was generated as a result of the discussions.

[17] https://www.sdsc.edu/services/hpc/tscc/condo_details.html

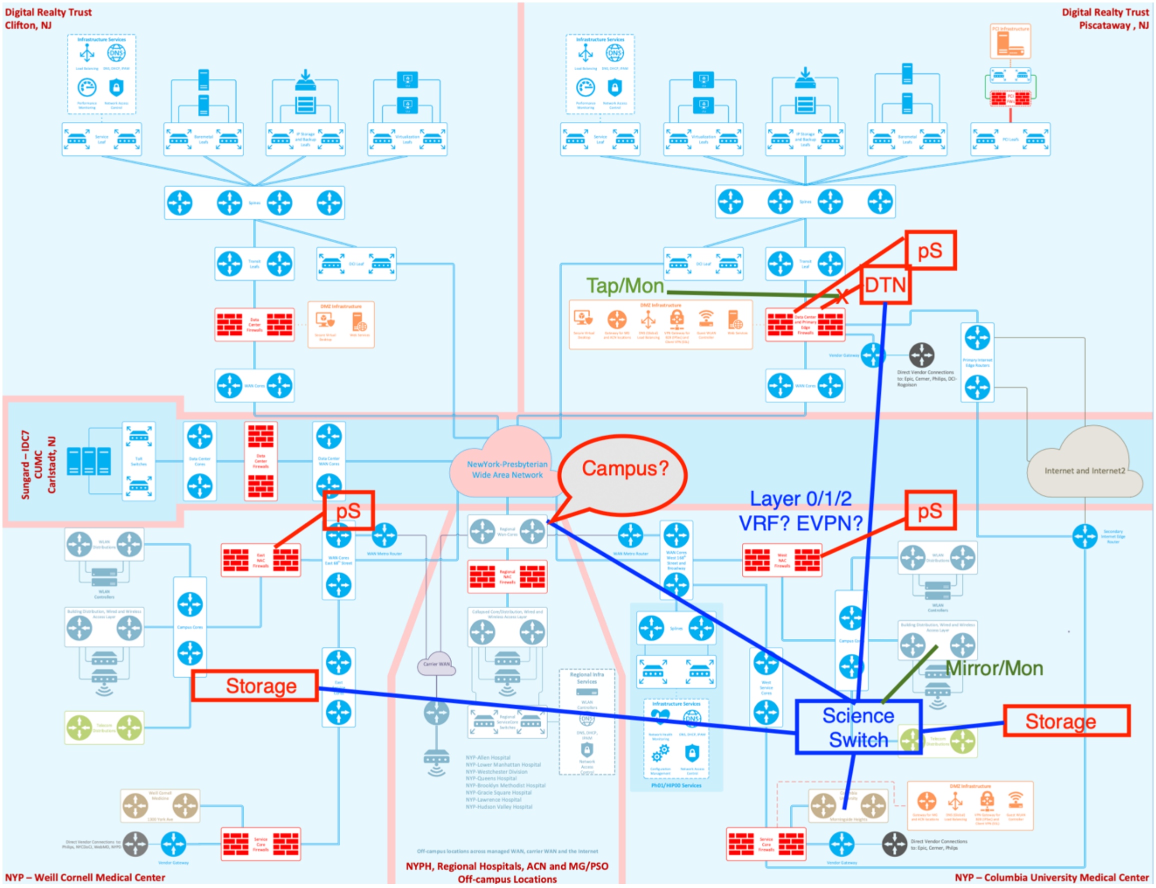


Figure 5.7.1 – Potential Clinical DMZ Environment

Some highlights to this discussion included:

- ***Creating pathways for data-intensive operations***: The first step of a data architecture is to identify where data must flow. In some cases, this may be "east and west" (e.g., within the campus), and in others it may be "north and south" (e.g., egressing or ingressing campus). From the use cases, there are a number of use cases that fit both of these definitions. An instrument that needs to store data to a SAN would fit the 1st model, a user looking to send information from the SAN to a cloud analysis platform would be the 2nd. To accomplish data mobility in either case, the technology group can leverage the network in the following way:
    - ***Existing network paths***: Traffic can be isolated once the path has been located. Using VLANs and QoS, it may be possible to offer different classes of services for data movement hardware in certain locations as it attempts to access storage or computation. Note this approach will rely on hardware that can honor traffic guarantees (may not be possible on older or less intelligent devices) and can become labor intensive to maintain configurations that are not automated.
    - ***New and/or dedicated network paths***: In a fiber-rich environment, creating new dedicated local paths from data producers to a new

"Clinical DMZ Switch" can facilitate a faster way to have data flow from instruments to storage resources.
    - Additionally, steps should be made taken to optimize network to provide high throughput to facilitate the sharing, storage, and analysis of data between intra-campus and cloud environments. Internet 2 capabilities and services could be utilized more effectively with additional, dedicated bandwidth connectivity, and tools such as DataSync would improve the ingress and egress abilities of data transfer into preferred storage locations.
- ***Use of overlay networks***: Like the discussion about on-campus network data mobility, creating overlay network relationships with other entities on the WAN may be beneficial when there are patterns of constant data transfer. For example, if a research group is routinely transferring data to a collaborator, working with that remote site to create a WAN-based VLAN (or VRF) may be beneficial. This is not a simple task, and does require coordination, but may reduce the burden security in the long run.
- ***Establishing application-level monitoring within and external to the networks***: perfSONAR testing, located at several places in and around campus, will provide a baseline network performance measurement to match against data mobility requirements. Testing to regional and national resources (e.g., NYSERNet, Internet2, ESnet), in addition to collaborators (e.g., JHU, UCSD) is recommended.
- ***Visibility into traffic with passive taps and analysis frameworks***: To prevent friction of data transfer, adopting a method of either mirroring switch ports, or using optical taps, can provide visibility into network traffic. This data can be fed into analysis platforms (e.g., Corelight[18]) that try to build understanding of network behavior and mitigate risk.
- ***Creating unified storage enclaves***: Storage on campus is a mixture of private solutions (e.g., local NAS, removable media) and some institutional capabilities (e.g., C2B2, etc.). Investing in institutional storage and making it available to instruments that require this capability, will ensure that research data has a unified ingress and egress. A campus-supported SAN will also allow researchers the option to purchase access and be assured that sensitive data is being handled appropriately.
- ***Common data mobility platform*:** Data transfer outside of the environment requires dedicated hardware and software. Creating a set of Data Transfer Nodes (DTNs) that have the capability to send 10Gbps+ streams using effective transfer tools (e.g., Globus[19]) will ensure that the storage enclave has a way to manage data mobility. Instances located around all campuses are recommended.
- ***Use case support, and proximity to firewalls***: positioning resources outside of the firewall is highly unlikely. If a DTN (or perfSONAR node) could be

[18] https://corelight.com/

[19] https://www.globus.org/

placed, the use of traffic inspection tools (e.g., taps, mirror ports) would be required.

- ***Common ways to egress the network that leverages the R&E and cloud connectivity options***: If a Clinical DMZ is established, it should have a way to leverage WAN connectivity directly. This can be a short hop to the border, or a direct connection to NYSERNET, Internet2, or clouds (e.g., Megaport).

Appendix A – NYSERNet

NYSERNet is a non-profit organization. Our mission is to advance the research and educational missions of our members by delivering a full range of customized, progressive, and affordable end-to-end data and networking technology solutions.

The networks we build with technology are only as effective as our community network. This is why we consistently cultivate opportunities to collaborate through our widely accessible professional development conferences and events.

NYSERNet began delivering network services long before most people heard of the Internet. In 1985, a group of visionary men and women from the state's leading research universities and institutions came together with the idea of creating a high-speed research network. Two years later, we deployed the nation's first statewide regional IP network.

In the 30 years since, NYSERNet has gone on to build a faster and more robust fiber optic network that offers colleges, museums, healthcare facilities, primary and secondary schools, and research institutions from Buffalo to New York City access to 100-gigabit speeds.

Today, we offer our members a fast network and data centers that help them take advantage of it. At our main offices in Syracuse, we have built a 4,000-square-foot facility that's key to our members' disaster recovery and business continuity strategies. In New York City, our Colo@32 is a place where national and international networks converge, providing unparalleled opportunities for connections.

At NYSERNet, education is a key part of our mission. Over the past 10 years, we've expanded our education program to offer our members hands-on technical training courses and an annual network technology conference called The NYSERNet Conference. Each year, we also bring together the state's most influential campus IT leaders for NYSCIO, the New York State Conference of Higher Education CIOs.