DEFENSE INNOVATION BOARD
JOINT PATHOLOGY CENTER (JPC) REPOSITORY ENHANCEMENT

Author: Eric Lander
Contributors: Kurt DelBene, Michael McQuade, Richard Murray
Staff Lead: Schuyler Moore

OVERVIEW

The Department of Defense (DoD) owns the largest repository of disease- and cancer-related medical data in the world. The Joint Pathology Center (JPC), part of the Defense Health Agency (DHA), is the premier pathology reference center for the federal government. It holds a tissue repository that includes approximately 55 million glass slides, 31 million paraffin-embedded tissue blocks, and over 500,000 wet tissue samples that have been collected over the last 100+ years. This repository represents a priceless resource for clinicians, pathologists, and health care data analysts to better understand and diagnose diseases, from infectious diseases to cancers. At the present time and in accordance with the National Defense Authorization Act of 2008,[1] the JPC is pursuing repository modernization efforts, to include converting the glass slide collection to digital slides.

The wealth of information housed in the JPC's repository has already been successfully leveraged in the past to address military and public health issues. For example, the tissue specimens in the repository were used to sequence the 1918 influenza virus, which killed more than 40 million people worldwide. The resulting research ultimately provided guidance for avoiding future influenza outbreaks that could affect military readiness, fighting strength and global health.[2] More recently, in 2017, tissue samples obtained from glass slides from the 1960's were used to demonstrate the presence of valuable vaccine targets for human melioidosis, a disease caused by an opportunistic pathogen widely distributed in soil and water in areas that Service members frequently deploy.

If the entire repository were leveraged to its fullest potential, it would advance diagnosis and treatment for thousands of illnesses, saving lives across DoD and the global population. A medical dataset of this size will rapidly accelerate biomedical research, particularly in the fields of infectious disease and cancer, reducing misdiagnosis by orders of magnitude and opening the door for more effective treatment options. In time, artificial intelligence and machine learning models may help pathologists sort through this massive dataset more quickly and effectively to provide better care for patients in and out of the military. It is difficult to overstate

[1] U.S. Congress, House & Senate, "National Defense Authorization Act 2008," 110th Congress, H.R. 4986, 122 Stat 200, https://www.congress.gov/110/plaws/publ181/PLAW-110publ181.pdf.
[2] Institute of Medicine 2012, *Future Uses of the Department of Defense Joint Pathology Center Biorepository*, Washington, DC: The National Academies Press. https://doi.org/10.17226/13443.

the importance and magnitude of this resource: no other pathology repository of this size and scope exists in the world.

A digitized JPC repository would accelerate progress in machine learning in medicine, with the added benefit of increasing DoD's standing in the field. Reference datasets have played a crucial role in advances in machine learning, by making it possible for researchers to test and compare the power of different approaches. DoD has great interest in machine learning. While DoD cannot share data related to national defense, creating a world leading medical-data resource would position DoD as a nexus of progress in the machine learning community.

Digitizing more of the JPC's repository would enable major improvements for the US population and for global healthcare. DoD in particular stands to benefit in several ways: (1) improving Service member health and readiness, (2) providing a resource to deepen existing partnerships and more effectively compete with peer adversaries, (3) providing a massive dataset to support DoD reform & modernization efforts in the field of AI/ML; and (4) lowering healthcare costs for DoD. These benefits listed above have a broad implication for the general population that faces many of the same health challenges as DoD. By appropriately leveraging this repository, DoD has the opportunity to be a prime contributor in the field of health care and improved medicine while improving its own operations and reducing costs.

These benefits, however, are currently lost because the vast majority of the JPC repository is still in a physical format, with few corresponding digital images to facilitate sharing and analysis. Moreover, the physical slides are also degrading over time. The JPC is currently limited by resources and infrastructure to digitize its full repository. Without further action, slides will continue to degrade and some may ultimately become damaged beyond repair. This loss, in combination with barriers to sharing and analysis, is handicapping a priceless DoD resource that could change the face of military health and readiness.

Critically, both physical and digital images of slides must be linked to their associated medical records to enable any truly substantive analysis of the repository. These medical records include annotations and metadata for the slide images, as well as the medical history of the patient, and will provide the critical context for the slide images that will enhance diagnostics for thousands of cancers and infectious diseases. Without these records and context, the slide images will have limited utility. The JPC currently links slide images to medical history using accession numbers unique to a patient and doctor's visit (e.g., John Smith's accession number is 1234567, and slides from his tissue sample provided on his third visit are labeled 1234567-03). While the current method is sufficient for small-scale analysis, future efforts will require unique identification numbers for each slide to more easily sort and analyze them at a large scale (using either a pathologist or AI/ML algorithms). Additionally, JPC slides are not currently tied to metadata in a standard format, as each scanner provides different types of metadata for images. In the future, standardized metadata collection will also be critical to enable large-scale analysis.

The JPC should explore approaches for adding molecular annotations to a subset of slides and tissue blocks to enhance their value for medical interpretation through machine learning. Molecular annotations include immunohistochemistry (IHC) to visualize the location of specific proteins, ranging from standard approaches involving a single antibody to multiplex approaches involving scores of antibodies simultaneously, and in situ hybridization (ISH) and spatial transcriptomics to visualize RNA expression. Molecular annotations illuminate rich detail about tissue that can then be scanned into digital format. Machine learning should make it possible to use molecular annotation from a subset of samples as the training date to enable the ability to *infer* such molecular information for the vast majority of slides, based solely on standard images.

For a visual representation of the JPC repository enhancement effort, see Appendix A.

DEFENSE INNOVATION BOARD RECOMMENDATIONS

In support of this tremendous potential to advance medicine and DoD's digital modernization effort, the Defense Innovation Board (DIB) held a biotechnology preparatory working session at the Broad Institute of MIT and Harvard in November 2019. DoD, academia, the commercial sector, and philanthropic leaders helped inform DIB subcommittee research. The DIB developed the following recommendations, based on the information gleaned from this session, interviews with subject matter experts and its own expertise. These recommendations provide DoD leaders with notional agendas and timelines that the Board feels are highly achievable: DoD should use these estimates as benchmarks for its own future agendas and timelines.

Recommendation 1: Pilot on Slide Scanning. The Secretary of Defense, through the Under Secretary for Personnel and Readiness (USD P&R), should direct a pilot scanning a large initial batch of slides as a foundation for implementing a longer-term plan to scan the full repository. This pilot should ideally start immediately and be completed within the next 12 months.

- The pilot project should pursue the following targets:
    - Volume: 1-2M slides
    - Timeframe: 12 months
    - Cost: ~$10-15M
    - Begin defining, adopting and scaling data management best practices for scanned slides as described in Recommendation 3
- The overarching goal of this pilot is to establish the methods, tools, processes and analyses that can then be improved and scaled so that slide throughput can be doubled in months 13-24 and the full repository can be scanned within a decade.
- After successful completion, it is critical that this pilot be rapidly scaled, and not cease or plateau at 2M slides/year. The JPC should ramp its scanning efforts, incorporating lessons learned to achieve economies of scale and set the pace to scan the full repository within the next decade.

Recommendation 2: Pilot on Linkage to Medical Records. While the JPC currently links slides to specific medical visits, slides do not have a unique identification number that would enable more rapid and more complete digital analysis, and many older records are not yet in digital format. The JPC should initiate a pilot to assess the ability to add value to the physical slide and tissue block collections by connecting a representative collection of slides to medical record information on a more automated and digital basis, to determine the extent of information and challenges. This pilot should ideally be completed within the next 12 months.

- The ability to connect samples to medical records is crucial to enable machine learning to identify features that predict specific diagnoses and outcomes. Analysis of 5,000 samples should suffice to determine the major issues.

Recommendation 3: Pilot on Molecular Annotation. JPC should initiate a pilot to assess the ability to add molecular annotations to physical slides and tissue blocks to enhance their value for machine learning. This pilot should ideally be completed within the next 12 months.

- Molecular annotations include immunohistochemistry (IHC) to visualize the location of specific proteins, ranging from standard approaches involving a single antibody to multiplex approaches involving scores of antibodies simultaneously, and in situ hybridization (ISH) and spatial transcriptomics to visualize RNA expression.
- Machine learning should make it possible to use molecular annotation for a subset of samples to *infer* such molecular information for the vast majority of slides, based solely on standard images.
- To evaluate recently developed multiplex approaches, various methods should be tested on a collection of 1,000 samples chosen based on medical interest.

Recommendation 4: Process Improvement Plan. DoD should begin a project to (i) develop a strategy to dramatically reduce the cost of slide scanning on an ongoing basis, with the initial goal of costs under $2 per slide (and ultimately much lower), to ensure that the effort is sustainable over the long term; (ii) develop a clear medical-records linkage strategy; and (iii) develop a clear data strategy. This project should ideally be started within the next 12 months.

- The JPC should **improve slide scanning** by:
    - targeting specific components of slide scanning to improve cost and speed while clearly defining a baseline of acceptable quality of outputs;
    - in some cases, improving individual steps (e.g., augmenting slide cleaning with computer vision or conducting bulk slide cleaning); and
    - in other cases, reimagining the whole process (e.g., building a separate scanning facility that incorporates automation and organizes slides to rapidly feed into the scanning process).
- The JPC, in coordination with other DoD stakeholders, should also **develop a clear data strategy** to promote effective data usage post-scanning. This strategy should set a roadmap for future JPC data storage and security while maximizing usability for

researchers and AI/ML models. A dataset of this size will require thoughtful data management planning, not only for storage and access logistics but also for data analytics usability. The JPC should work with the Defense Digital Service (DDS) to develop this data strategy, leveraging DDS' considerable experience in the field of data architecture and management.

- Given the anticipated sharing and access requirements of the JPC, the scanned slides will likely need to be organized into a data lake with an accompanying data catalog to promote data visibility and accessibility, ideally built in a cloud platform. The data should be stripped of explicit patient identifiers ("de-identified") and shared with DoD-approved partners.
- As the repository dataset enables more robust and accurate AI/ML models, those models may also be published in the data catalog as a shareable resource for researchers.
- As the JPC works to continuously improve its slide scanning and data management processes, each iterative approach can be used as a data tag to track the progression of improvement and more easily identify successful (and unsuccessful) approaches.
- Data management suggestions are included in Appendix B.

Recommendation 5: Partnership Plan. DoD and the JPC should develop, ideally within the next 90 days, a Partnership Plan that clearly defines desired partner archetypes (e.g., pathology expertise, AI/ML expertise, data storage expertise) and partnership structures.

- The JPC will need to build partnerships across research, commercial, and government sectors to facilitate rapid scanning of the repository and promote productive use of the digitized repository post-scanning. Specifically, the JPC should reach out to organizations and institutions who have conducted similar exercises, such as the Memorial Sloan Kettering Cancer Center (MSKCC), to learn slide scanning best practices and collaborate on future scanning efforts. (See Appendix B for details about the MSKCC slide scanning efforts.)
- The JPC will ultimately need to build a consortium of partners providing diverse expertise, equipment, and funding in order to scan the full repository.
- This plan should seek to share this resource as widely as possible across the research community. Examples like the Human Genome Project and other genetic datasets demonstrate the public good that can be derived from providing cost-free access to massive medical datasets (while appropriately protecting medical information and patient privacy). For non-profit research users, access should be cost-free. For commercial users, any costs should be designed to ensure that the repository is widely accessible, including to start-up companies.

Recommendation 6: Ultimate Goal - Fully Digitized and Accessible Repository. To reach the full potential of the repository and based on the results of the pilot projects, DoD should aim to ensure that all physical glass slides have been replicated in digital format with associated relevant medical records, with ethical and secure handling of personal health information (PHI) and a data lake (DoD-owned IP) accessible to approved parties. This transformation should occur within 5-10 years after the completion of the pilot projects.

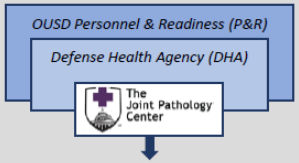
- In pursuing this objective, **DoD and the JPC must bear in mind two critical factors**:
    - *Scale:* The current approach, infrastructure, and personnel applied to scanning small pockets of the repository is insufficient. To transform the repository, the JPC must develop its approach, infrastructure, and personnel with scale in mind. Current processes for scanning a few thousand slides will not be appropriate for scanning 55 million slides. The JPC must plan accordingly to scan the full repository at a reduced cost per slide.
    - *Time:* The physical repository degrades every day, with lives lost that might have been saved with the information in the repository. The JPC should move quickly to save and leverage this resource. The "perfect" should not be the enemy of "good enough" - the JPC should scan large batches of slides now and work through challenges in a continuous manner, rather than wait until it has the "perfect" execution plan.

DIGITIZING THE JOINT PATHOLOGY CENTER:

Harnessing medical data to enhance DoD on the physical and digital battlefield

CURRENT STATE

DoD owns the **largest repository of disease- and cancer-related medical data in the world.**



55 million glass slides
31 million paraffin-embedded tissue blocks
500,000 wet tissue sample
Collected over 100+ years

This resource has the ability to **create step-change improvements in DoD and global health care** when fully leveraged.

BUT…

- Repository is **hard to share** (physical format)
- Repository is **hard to analyze** (lack of digital data)
- Repository is **degrading over time** (aging slides)

TRANSFORMATION

To better leverage the repository, the JPC must convert its glass slides into a more **shareable & useable digital format AT SCALE.**

Mechanics: Identify and enhance steps of slide scanning "mechanics" that have an outsized impact on a) cost and speed of scanning and b) quality of outputs

Data management: Leverage commercial solutions to structure data storage and security while maximizing usability for researchers and AI/ML algorithms

Business model: Build partnerships across research, commercial sector, and government to facilitate rapid advances in public health while building in cost-sharing mechanisms to sustain the digital repository long-term

TARGET END STATE

The Joint Pathology Center

DIGITAL

Physical glass slide repository replicated in digital format

SECURE

Ethical/secure handling of personal health information (PHI)

ACCESSIBLE

Data lake (DoD-owned IP) accessible to approved parties

IMPACT ON DoD

READINESS

Increased health & readiness boosting deployment, recruitment, and retention

ALLIES & PARTNERS

Deepened data partnerships, more effective competition with peer adversaries

REFORM & MODERNIZATION

Reduced DoD health care costs and improved digital modernization in the field of data management and AI/ML

*To accomplish this transformation, DoD should pursue a **scalable JPC pilot to scan a large initial batch of slides in the near-term** (e.g., 1-2M slides scanned within the next 12 months, ~$10-15M).*

*Using those lessons-learned, the JPC can then **scale and implement a longer-term plan to scan the full repository.***

APPENDIX B

For the purpose of this report, respective areas of potential improvement and partnership are divided into two focus areas: slide scanning mechanics and slide scanning data management. These areas represent lines of effort critical to the digitization and management of the JPC's repository. To digitize the JPC repository **at scale, at speed of relevance, and at reasonable cost**, the JPC will need to carefully consider its current process to improve on certain steps and entirely reform others as necessary.

Slide Scanning: Mechanics

The current process for slide scanning is as follows, using a sample set of 100 slides:[3]

1. Slide selection/prioritization
2. Slide retrieval from repository for scanning
3. Slide cleaning
4. Slide data linkage
5. Slide scanning
6. Slide image quality control post-scan

The current approach to scanning slides at the JPC would take a century or more to digitize the full repository. There are a range of traditional and non-traditional methods for the JPC to expedite this process: the most viable path forward may require a mix of both, depending on the resources and political capital available.

For each of the five steps in the slide scanning process, short- and long-term opportunities to improve them are available via innovative technology, process refinement, and/or policy adjustment.

1. Slide selection/prioritization

The digitization effort would benefit from a thoughtful strategy regarding order of scanning (e.g., grouping and prioritization of slide batches). Previous efforts have selected smaller batches of slides related to a specific disease or cancer in order to target that disease or cancer for research. However, the digitization of the full repository may require an approach to slide selection that is agnostic of content and instead focuses on the slides that can be scanned most rapidly. For example, newer slides frequently require less cleaning than older slides that have been sitting in the repository for decades. Older slides may demand more individual repair that would disrupt the initial workflow of slide scanning: while preservation of these slides is critical, it may be appropriate to wait to scan older slides until the JPC has achieved high throughput of new slides.

[3] Slide scanning process and detailed descriptions drawn from interviews of JPC staff and other experts in the fields of pathology, slide scanning, and AI/ML for medical data.

If the JPC were to prioritize batches of new slides first to scan and then apply those lessons-learned to older, more difficult-to-scan slides, this might help expedite the process and reduce cost in the long-term. Within these batches, it may then be logical to group slides by disease or cancer, but the strategy must be devised with the full repository scanning in mind. If the strategy does not reflect the intention to scan the full repository, then the JPC will not be able to achieve economies of scale and will build in unsustainable time and cost to the plan.

For the first step in the process, the JPC must carefully choose a strategy for slide selection and prioritization when scanning the full repository to determine the order in which slides will be scanned. There are a number of ways in which this selection can occur, such as by disease/cancer type or by age of slide. In previous scanning efforts, the JPC has typically prioritized slides for scanning focused on a subset of disease or cancer types. This approach was valid in the context of narrow scanning objectives, but future efforts to scan the entire repository must take into account economies of scale: the infrastructure, manpower, and resources required for such a large undertaking will not be reached if the scanning plan only accounts for small batch scanning.

Scanning the full JPC repository calls for a fundamental reset in approach, given the size and breadth of its content. Part of this approach will require scaling existing systems and processes (e.g., buying more scanners and hiring more personnel), but others may require a paradigm shift in approach by integrating alternative methodologies, such as automation or computer vision. However, the value of these alternative methodologies only becomes clear in the context of the entire repository. If future plans for the JPC fail to think long-term about scanning at scale, those plans may be doomed to include unnecessary time and cost burdens.

2. Slide retrieval from repository for scanning

The second step of physically retrieving slides to scan adds time to the scanning process, given the size of the repository storage vaults. When selecting sets of slides for scanning, pathologists and technicians must locate and pull slides out of the thousands of boxes currently stored at the JPC. While the JPC's laboratory information system can help direct technicians to the location of the slides, the process of physically locating and retrieving slides is highly manual and creates significant drag of the process. If the JPC begins to scan its repository at scale, it may need to consider alternative facility design to better facilitate the scanning process, including physical storage reorganization and automated retrieval.

The process of retrieving slides from the repository relies on the laboratory information system as a map of slide location but is ultimately manual, requiring heavy human interaction with slides due to the size of the collection and lack of automated retrieval. Once a clear slide prioritization strategy has been established and the JPC has identified its first tranche of slides to scan, it may need to consider options for slide storage and organization for those slides that reflect that strategy and order of slide scanning by physically arranging slides to optimize retrieval. Each

tranche of slides should serve as a test case for improving slide retrieval, either through automation, more efficient human process, or otherwise. Each tranche should seek to improve on the time and cost of the previous ones, improving the process over time while continuing to make progress on the overall slide digitization objective.

3. Slide cleaning

The third step of slide cleaning requires close visual scrutiny and cleaning by a trained technician before being scanned. This step is necessary to produce clear, readable slide scans for future users to digest, and takes varying amounts of time depending on the age and condition of the slide. While more recent slides may require a simple scrub with cloth and cleaning solution to remove fingerprints and other surface blemishes, older slides have dried out or crystallized to the point where the slide must be re-cut or manually scanned. The current process is labor intensive, and adds significant time and cost to the scanning process.

Slide cleaning represents one of the more time- and cost-intensive elements of the scanning process, as it is highly manual and requires close visual inspection of each slide to determine whether and to what extent it needs to be cleaned. For this reason, the JPC should consider ways in which it can reduce the time spent on cleaning, either for individual slides or in aggregate. For individual slides, it may be possible to augment the current cleaning process through automation or more rapidly identify slides that require cleaning using computer vision and other AI/ML models to improve process efficiency. In aggregate, there may be an opportunity to clean continuously or in bulk to create a steady stream of slides that are ready to be scanned. Increased manpower and/or automation could allow the JPC to clean slides at a consistent rate, limiting the bottleneck between slide retrieval and slide scanning. Alternatively, the JPC could consider bulk cleaning methods that could address speed of cleaning by targeting large batches simultaneously, which in turn may reduce the requirement for increased manpower that might be required for continuous scanning using the current method.

4. Slide data linkage

Cleaned slides must then be linked to their relevant paperwork and metadata. This process leverages the JPC's existing laboratory information system, which holds detailed information about each slide's contents as well as the medical context of the individual from which the tissue was drawn. This data linkage is critical for future use of the digitized repository. While the slide images alone carry value for cancer and disease research, the associated contextual data will define the true value of the repository as a whole if it is leveraged to identify broader trends and anomalies that have previously been obscured due to lack of data. While some research organizations like the Memorial Sloan Kettering Cancer Center (MSKCC) have begun to link metadata more cleanly to slides through barcoding, the JPC does not currently have those linkages built in. As a result, the current format and process risks devaluing the resulting digitized repository if slide images cannot be readily linked to their associated contextual data.

The JPC's current data linkage approach is highly manual, connecting physical slides to their associated contextual and metadata. The JPC is beginning to implement barcoding to new slides in the repository, but this does not address the 55 million slides that are currently without barcodes in the repository. The JPC may want to consider partnering with other institutions facing similar challenges (such as MSK) to find more efficient methods of tying data to slides without barcodes.

5. Slide scanning

The fourth step of the process, scanning of the cleaned slides, is highly dependent on the type of scanner used. Different scanner brands and models have different slide capacity, speed, resolution, pixelation, and magnification capabilities. On average, modern scanners can hold ~100-300 slides at one time, and take ~1 minute per slide (~2-5 hours to scan all slides in a full scanner, depending on its loading capacity). This part of the process is clearly time and cost intensive, but is highly dependent on advancements by slide scanner manufacturers. Scanner throughput is increasing over time, but there are both technical and policy barriers to advancement in this field. From a technical perspective, scanners are slowly evolving to allow for faster, clearer scanning. From a policy perspective, scanner manufacturers are currently limited by the fact that the FDA has typically classified their products as Class III medical devices, bringing with it a host of regulatory restrictions that hamper manufacturers from updating and releasing new scanners onto the market more rapidly. Only recently has the FDA adjusted its stance to classify some scanners as Class II medical devices, but this process is slow-moving and restricts rapid upgrading and deployment of better, faster scanners.

The cost and speed of slide scanning is dependent on the speed of slide loading and the speed and quality of the scanner itself. For slide loading, there may be room for improvement in speed of loading via automation or altered design of slide scanners to better complement the manual process of loading. The importance of improving efficiency of this step may fluctuate depending on the future development of scanners: for example, if scanners develop to hold more slides (greater than the current capacity of ~100-300), then the time required to load a full scanner may become prohibitive and may warrant closer examination. However, if scanners maintain a steady loading capacity and instead speed up the scanning process, this step may require less examination.

Regarding the speed and quality of the scanner itself, current models vary in slide scan image quality and throughput. While some scanner models may have better image quality, this does not necessarily come with improved logistics and increased scanning throughput. Organizations may have to choose between the two, and the value of either may depend on the organization and intended use of the slides. For example, in the case of the JPC it may be worth increasing throughput if the sacrifice in image quality is not so severe as to limit use of the images. The JPC has limited ability to guide development of future scanners, but it can thoughtfully and practically apply those scanners. For example, it may be more effective to lease some scanners

than to buy them outright, allowing for future upgrades as new models are released and less sunk cost once the repository has been scanned and scanners are only required for a small number of new incoming slides. The JPC will need to devise a strategy for scanner acquisition and usage that accounts for future repository requirements and data usage.

6. Slide image quality control post-scan

The fifth and final step of the process involves quality control vetting of images to ensure that the scanned images are usable. Like the cleaning process, this requires human review, and conducting quality control for a batch of ~100 slides can take hours. New generations of scanners have image quality control built into their process so that they can flag distorted images and automatically rescan those slides, but this functionality is imperfect and frequently still requires human review. This step is critical to building an effective and useful repository database, and also offers an opportunity to significantly reduce the time of scanning.

In the long term, the JPC can and should use the quality control process to improve its scanning process. The JPC may be able to apply computer vision algorithms on newly-scanned images to detect patterns of defects or imperfections in images, which can then be retroactively adjusted for. This will reduce the time required to re-scan certain slides, allowing for faster scanning overall. From a policy perspective, upgraded scanners may become more readily available if the FDA continues to move scanners down from class III to class II medical devices, removing significant hurdles for manufacturers to take their products to market and increasing the range and quality of products available to the JPC.

Case Study: Memorial Sloan Kettering Cancer Center[4]

The Memorial Sloan Kettering Cancer Center (MSKCC) can serve as a relevant proxy to the challenges and opportunities faced by the JPC. MSKCC's Department of Pathology is one of the largest departments in the world, with a massive archive of over 25 million glass slides and pathology information dating back more than 40 years.

MSKCC has recognized that "inherent in the use of glass slides are the logistical issues of slide transport, archiving, and retrieval, as well as limitations in remote review, conferencing, and consultation capabilities — challenges most pathology laboratories face today. Pathology as a discipline is also stymied by increased workloads and a shortage of specialists in some parts of the world. There is a significant need for tools that can reduce the burden and increase the precision of cancer diagnosis."[5]

[4] Sourced from interviews of MSKCC staff and program leaders.
[5] "The Warren Alpert Center for Digital and Computational Pathology at MSK," MSK, accessed 01 February 2020, https://www.mskcc.org/departments/pathology/warren-alpert-center-digital-and-computational-pathology.

Improvements in technology and recent FDA approvals for upgraded scanning machines have created the opportunity to overcome some of these barriers. The overlap between MSKCC's expertise in pathology, digital imaging, and large-scale machine learning in a data-rich environment has allowed MSKCC to begin scanning its repository at scale. In 2012, MSKCC began barcoding the ~1 million slides produced annually at its facilities, and then scanning those barcoded slides. In the ensuing years, MSKCC has ramped its scanning operation and scanned nearly 1 million slides in 2019 alone, with the target of scanning a total of 4 million slides by 2020. MSKCC has managed to scale its scanning operation by setting up an offsite location scanning operation for archive slides, staffed by a robust team of full-time equivalents (FTE) scanning slides at a rapid clip. MSKCC's scanning process has been streamlined to account for the scale of the operation, taking the steps outlined earlier in this paper and conducting them in bulk or prioritizing rapid scanning in combination with post-scanning reconciliation to facilitate increased throughput of slides. Additionally, many of MSKCC's scanners are leased, enabling them to more rapidly upgrade machines as new variants come out and reducing sunk cost once the bulk of archive scanning is completed.

To facilitate computational pathology in practice, the MSKCC team is building on a fully digital workflow and advanced imaging techniques in digital pathology. Recent breakthroughs in machine learning and improvements in scanning technologies have enabled MSKCC to build decision-support systems that use deep learning and artificial intelligence at an unprecedented scale. Between slide process mechanics and data management, the MSKCC represents a highly relevant case study for the JPC to observe and absorb lessons-learned.

Slide Scanning: Data Management

While scanning the full repository will require thoughtful planning from a scanning-mechanics standpoint, it will equally require thoughtful planning for management of the data post-scanning. This planning process can be described in three parts:

1. Data storage
2. Data security & access
3. Data analytics

Each of these steps has implications for the slide scanning mechanics described earlier. Data storage and security & access each require appropriate tagging of physical slides and digital images to relevant patient data and metadata in order to store and protect private information appropriately. Data analytics require even more attention to detail for data tagging and linkage, as those characteristics will provide the baseline for any traditional or AI/ML- driven analytics.

1. Data storage

The JPC will need a hardware and software acquisition plan to accommodate the vast amount of data rendered from the repository scanning process. Various commercial solutions already exist to provide data storage - industries ranging satellite imagery to banking have required data storage at massive scale for years, and demonstrate the current availability of solutions that the JPC can leverage. Finding solution options may be a straightforward process, but finding a solution at a manageable cost level will be more challenging. Scanned slide images file sizes are typically between 1-2 gigabytes[6]: this means that the full repository of 55 million slides represents a data source on the order of magnitude of exabytes. Given this size, data storage will contribute significantly to the cost of the effort (although it will largely come in the form of a single "up front" cost that will not be repeated in ensuing years). Additionally, the JPC should build in storage redundancy to ensure continuous data availability and reduce risk of losing data.

To manage cost of data storage, the JPC will need to consider the various commercial storage options available, ranging from in-house storage to cloud, and select an option based on current and projected data use bearing in mind the scale required for a repository of this size. Additionally, a dataset of this size will require a thoughtful strategy for where to run compute. To promote security of the dataset, the JPC should ultimately aim to couple compute with the data lake to limit the need for copying data. This approach will also allow for better large-scale analysis of the repository, as copying data at scale to develop AI/ML models may be impractical.

2. Data security & access

Once the JPC has determined its data storage strategy, the next challenge will involve security of and access to that stored data. Security is particularly critical in the case of the JPC, given the wealth of personal data that will result from the slide scanning process, and access to the data will inherently be correlated to the issue of security. The JPC is not unique in being a caretaker of sensitive personal data: other medical organizations, government organizations, financial institutions, and others bear the same responsibility and have security products that defend that data accordingly. There are a broad range of data architecture strategies and commercial products that can ensure the security of the JPC's data, as well as control access and retrieval of that data. There are a variety of "data de-identification" products that separate Personally Identifiable Information (PII) from Protected Health Information (PHI), creating a category of data that cannot be tied back to an individual while maintaining the richness of the JPC's data. This process also provides more storage flexibility, as de-identified data does not fall under HIPAA.

Once data has been de-identified, the JPC will need to consider the range of potential partners that can benefit from access to this resource, and structure the data accordingly to allow a broad

[6] Yukako Yagi et al., "An Ultra-High Speed Whole Slide Image Viewing System," Analytical Cellular Pathology (Amsterdam) 35 (1): 65-73, 2012, https://dash.harvard.edu/bitstream/handle/1/23993570/4605572.pdf?sequence=1&isAllowed=y

set of researchers to use the data to its fullest potential. The JPC can filter data access and use based on security concerns, but can also incentivize research focus areas by providing access for health issues critical to DoD, and promote collaboration with non-governmental organizations by offering more access to organizations partnering with DoD. The JPC repository is a priceless resource for the medical research community, and provides an opportunity for DoD to strengthen connections to that community while promoting military readiness and health across the force.

3. Data analytics

Once the slides have been scanned, the data stored, and security and access controls set, the JPC and the broader research community can realize the true value of the repository by leveraging the massive dataset to find previously-undiscovered trends and anomalies related to infectious disease and cancer. However, the volume of data may be prohibitive for human researchers to ingest: instead, those researchers may need to rely on computer-assisted viewing and analysis to more effectively use the repository. Artificial intelligence (AI) and machine learning (ML) models are increasingly being used in the medical field and may create step-change improvements in infectious disease and cancer diagnostics. Potential AI/ML use cases and benefits are described in the Digital Pathology Association's whitepaper, "A Practical Guide to Whole Slide Imaging:"

> "Pathology-centric ML approaches include cellular heterogeneity and stromal feature extraction algorithms, which have proven to be quite useful for prognosis in the setting of breast carcinoma. Similarly, ML approaches to analyze tissue specific morphologic features have also demonstrated utility for prognosis in the setting of lung carcinoma. More recently, deep ML methods like convolutional neural networks have become more popular in the biomedical setting. These represent a fusion of traditional computer vision approaches with modern ML optimization, where the computer selects both the intermediate features that are extracted and the learning applied to those features within a single model. Thus far, deep ML techniques have been used for image segmentation, object classification, object recognition, and clinical outcomes prediction. More specifically, in the setting of pathology, deep ML methods have been extensively researched and applied to H&E-stained whole slide images for tumor region identification, detection of metastatic foci, tumor classification, and prediction of gene mutations. Standardized frameworks exist that, for a suitably annotated body of reference data, will meet or exceed in an automated fashion the performance of any one pathologist. As such, artificial intelligence techniques, and artificial intelligence–derived methods including computer vision, ML, and deep ML, promise to provide pathologists with a number of useful tools, beginning with mechanisms for automated case review and eventually leading to computer-aided diagnosis. These tools, which will undoubtedly

enhance pathology workflows, will ultimately play a larger role in improving patient outcomes."[7]

As demonstrated above, AI/ML can enhance the field of pathology, providing pathologists with time-saving mechanisms to more rapidly and accurately analyze tissue slides. As with any AI/ML algorithm, pathology-focused algorithms will require reams of tagged data to train and refine themselves, and the JPC repository could provide a unique opportunity and environment for DoD to improve its algorithms while improving health care across the Department and beyond. As these algorithms improve over time, they may contribute to the slide annotation process by rapidly identifying areas of the slide on which pathologists should focus, identifying relevant "clusters" of image content and cutting down on human analysis time. AI/ML will assist with the logistics leading up to analytics and the analytics themselves, making better use of the digitized repository and providing significant contributions to the research community.

AI/ML-assisted data analytics for the repository will require close linkage between the scanned slide images and the associated metadata to enable supervised and unsupervised learning. This means that the JPC will need to establish best practices for tying in metadata early in the scanning process: for new slides sent to the repository, the JPC should consider barcoding techniques currently deployed in institutes like MSK, while old slides may require a mix of barcoding or metadata tagging post-scan. These practices will also help provide DoD with a training ground to improve its algorithms for a universally positive test case.

Case Study: IEEE ISBI Challenges

Biomedical imaging challenges can provide examples of the successful development of deep learning approaches to pathology sample analysis as well as lessons learning for organizing such training libraries.

The field of image processing, more specifically computational pathology, is rapidly evolving with the best algorithms already functioning at a comparable level to professional pathologists. International challenges, such as the IEEE International Symposium on Biomedical Imaging's (ISBI) grand challenges, have provided an opportunity for leading innovators in academia and industry to compete and refine computational methods. The IEEE ISBI 2016 Camelyon Grand Challenge (CAMELYON16) pitted 11 pathologists against deep learning algorithms in the detection of lymph node metastases in whole slide images. After training on a moderately sized labeled library (n=110 samples with metastasis; n=160 without metastases), the top-performing algorithm performed significantly better than the pathologists in the whole slide imaging

[7] Mark Zarella, et al, "A Practical Guide to Whole Slide Imaging," Digital Pathology Association, Archives of Pathology & Laboratory Medicine, Vol 143, Feb 2019, https://www.archivesofpathology.org/doi/pdf/10.5858/arpa.2018-0343-RA.

classification task.[8] A joint team from Harvard Medical School and the Massachusetts Institute of Technology developed this winning algorithm, which employed image pre-processing to exclude background space and reduce computation time, followed by a cancer metastasis detection framework including a patch-based classification stage and a heat map-based post processing stage. Four well-known deep learning network architectures, GoogLeNet, AlexNet, VGG16, and a face oriented deep network were evaluated for the patch-based classification, with GoogLeNet performing at the highest accuracy. The model then produced a tumor heat map and classified test cases at a greater than 90% accuracy.[9]

More recently, the 400 stained slides from the CAMELYON16 dataset were used to produce PatchCamelyon (PCam), a library of over 300,000 patch samples derived from whole slide images. PCam provides a new benchmark for machine learning models and has been used to show the effectiveness of rotational equivariant convolutional neural networks in analyzing pathology data.[10] Designing for flexibility in post processing of the digital whole slide image library may enable the application of novel machine learning techniques.

As evidenced by CAMELYON16 and PCam, deep learning networks have demonstrated the capability to classify pathological samples as accurately as human operators for years, but only with a sufficiently annotated training library. Adequate slide annotation requires both expert attention and time. Due to resource constraints, digitization of the JPC should include slide annotation as a parallel process to slide-scanning to prevent a process bottleneck. Computer aided processes for filtering background and highlighting regions of interest may aid in expediting the annotating process.

Various machine learning techniques may also require different forms of training libraries as input. Separating the digital library of annotated whole slide images from its derived training sets as stack and flow may prevent inadvertent data corruption at the cost of storage space. In other words, the JPC would do well to prioritize developing a protected, digital library with detailed annotations for use in the development of various training sets. The rapid evolution of machine learning techniques for image processing, specifically in digital pathology, suggests the development of a flexible library is more important than designing a library for use with any single machine learning model in mind.

---

[8] Ehteshami Bejnordi et al., "Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer," *JAMA,* 318, no. 22 (12 Dec 2017): 2199–2210, https://doi.org/10.1001/jama.2017.14585.

[9] Dayong Wang et al., "Deep Learning for Identifying Metastatic Breast Cancer," *ArXiv:1606.05718 [Cs, q-Bio]*, 18 June 2016, http://arxiv.org/abs/1606.05718.

[10] Bastiaan S. Veeling et al., "Rotation Equivariant CNNs for Digital Pathology," *ArXiv:1806.03962 [Cs, Stat]*, 8 June 2018, http://arxiv.org/abs/1806.03962.