Pathology Archive

Evaluation of Integrity, Regulatory Compliance, and Construction of Searchable Database From Print Reports

Matthew A. Smith, MD, E. Leon Barnes, MD, and Simion I. Chiosea, MD

Key Words: Archive; Quality; Scanning

Abstract

Tissue repositories maintained by pathology departments represent an abundant resource of clinically annotated human specimens. The storage expenses associated with pathology archives are known to administrators of most pathology departments. However, such basic repository characteristics as the quality of stored materials, ease of access, and search and retrieval rates are often unclear.

The aims of our work were to design a framework to assess the quality of a historic pathology archive, to propose the definition of “archive integrity,” and to provide benchmarks for tissue block retrieval rates and DNA integrity. We share our experience with scanning approximately 120,000 pathology reports from 1956 to 1979 into an electronically searchable archive, with a $9,000 budget, completed in 6 weeks. Several ethical and legal considerations that shaped the technical side of this project are discussed.

To study epidemiologic and molecular aspects of disease over time, one needs a reliable source of material. Tissue repositories maintained by pathology departments represent an abundant resource of clinically annotated human specimens. The research value of pathology archives is highlighted by the National Cancer Institute–funded Shared Pathology Informatics Network, an effort to facilitate investigation using routinely stored formalin-fixed, paraffin-embedded tissue blocks.

The storage expenses associated with pathology archives are known to administrators of most pathology departments. However, such basic repository characteristics as the quality of stored materials, ease of access, and search and retrieval rates are often unclear. Assessment of these archive characteristics is financially prudent because it may help to save on storage expenses. For example, inadequate archive components can be discarded without harm to the department’s research mission. With the work presented herein, we pursued 2 goals. Our first aim was to design a framework to assess the quality of a historic pathology archive. Our second aim was to share our experience with scanning approximately 120,000 pathology reports from 1956 to 1979 into an electronically searchable archive at a budget of about $9,000, which was completed in 6 weeks.

Materials and Methods

Historic Pathology Archive

At the University of Pittsburgh Medical Center (UPMC), Pittsburgh, PA, pathology records from 1981 until the present are digitized. Anatomic pathology records before 1981 remain
on paper and will be referred to as the “historic archive.” In this context, “historic” refers to the components of the pathology archive that lack clinical relevance. The existing requirement by the College of American Pathologists mandates storage of pathology reports, slides, and tissue blocks for a minimum of 10 years. Currently, some patients with cancer survive for more than 10 years, and review of previous pathology material may be required for comparison with recurrent tumor or for enrollment in clinical trials. Therefore, the UPMC Department of Pathology currently maintains its pathology archive for at least 50 years. Most likely, adult oncolgic patients are deceased 50 years after the initial surgery, making corresponding pathology material “clinically irrelevant.”

The entire archive (paper records, glass slides, and tissue blocks) is professionally stored off-site in a climate-controlled environment, and its index is available on-line (Iron Mountain, Boston, MA). This Web page is maintained by the department of pathology and Iron Mountain. While the index has records of everything stored, its content is not organized chronologically. Also, alternative designations of pathology reports included “pathology books,” “laboratory books,” “volumes,” and “records.” To assess the chronicologic continuity of print reports, all of the aforementioned terms were searched, and a complete list of all available reports was created. Blocks and slides are requested online and delivered the next day. However, for material before 1981, the only available data-mining method was manual review of the paper records.

The physical condition and accuracy of the archive were initially assessed by direct manual search and review of 30 rare and common tumors from 1 volume of bound reports from 1956. An archived surgical pathology case was considered accurate when the accession number on the report, tissue block, and glass slides matched. Assessment of the tissue blocks included examination of the H&E-stained recuts.

**Legal and Ethical Clearance**

In consultation with the institutional review board, it was decided that as long as the archive is not extensively searched with a specific research goal, conversion of the paper archive into a digital format constitutes a change in the physical condition and is not categorized as “research.”

**Scanning of Paper Reports**

Paper records were stored as bound volumes. All documents were 210 × 280 mm. To prepare reports for scanning, the binders were removed using a rotary band saw. A cut was made through the pages, near the spine. Private scanning businesses in the Pittsburgh area were surveyed for compliance with private health information regulations. All scanning was performed by Compucom (Pittsburgh, PA) using a high-throughput scanner with a 500-sheet automatic document feeder and a dedicated personal computer workstation with image capture software and optical character recognition (OCR) installed on the premises of the department. Each scanned volume of reports was saved as an individual portable document format (pdf) file. The actual scanning (with removal of radiographs of gross specimens and electron microscopy photos) was performed by the company’s technical personnel. Once the reports were saved, the files were indexed and catalogued in Adobe Acrobat Professional 7.0 (Adobe, San Jose, CA).

To test DNA integrity, fluorescence in situ hybridization (FISH) for *MECT1/MAML2* was performed on archived mucoepidermoid carcinomas (MECs), as previously described.²

**Results**

**Assessment of Archive Quality: Reports, Slides, and Blocks**

The research value of an archive is directly related to its integrity. First, we determined the year of archive integrity, the earliest year when all 3 components of the archive (print reports, tissue blocks, and glass slides) were available (1956) [Table 1].

Next, bound volumes of print reports from 1956 were requested and received in excellent physical condition (Image 1A). Details of working with archived tissue blocks and glass slides are summarized in [Image 2a] and [Image 3a]. All received blocks were in good physical condition, and H&E-stained recuts were acceptable for interpretation.

The DNA integrity in cases from 1956 was tested by FISH for the *MECT1/MAML2* translocation on cases of MEC. Because the test was technically successful on 3 cases from the late 1950s, it was performed on 17 additional cases of MEC diagnosed from 1956 to 1974. FISH was technically successful in 14 of 20 cases.² Cases with failed tests (6/20 [30%]) were randomly distributed throughout the 18 years. It was decided that 1956 is the earliest year when all archive components are of an acceptable quality for research.

The knowledge of the earliest year of archive integrity is needed to prepare a detailed “business case” for archive scanning, ie, to calculate the number of pages to be digitally converted. To determine the number of reports, we reviewed the last volume of each year from 1956 to 1979 [Figure 1a]. Given the acceptable quality of the archive, the rate-limiting factor in its utilization was the lack of an automated way to search it.

**Ethical and Legal Considerations**

Before proceeding with construction of a searchable digital archive, several ethical and legal considerations that shaped the technical side of this project must be mentioned [Table 2].
Even though the pathology archive from 1956 to the 1970s predates the Health Insurance Portability and Accountability Act (HIPAA), it was recommended that these records should be treated as private health information, and the company that would perform scanning should agree to UPMC’s HIPAA business associate conditions. The practical implications of these decisions were as follows: First, bound volumes required removal of the book spine (Image 1B) to allow high-speed scanning. While removal of the book spine can be done off-site, concerns about patients’ confidentiality while transporting and storing records with a specialized contractor would require approval from regulatory departments.

| Table 1 Framework for Assessing Archive Integrity*

<table>
<thead>
<tr>
<th>Archive Component</th>
<th>Criteria</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Print reports</td>
<td>Identify the earliest year when print reports are in good physical condition, are chronologically continuous (cover most of the year without significant span skips), can be retrieved at acceptable rate, and can be prepared for potential high-throughput scanning (Image 1)</td>
<td>In our department, mid-1940s</td>
</tr>
<tr>
<td>Tissue blocks and glass slides</td>
<td>Identify the earliest year when tissue blocks and glass slides are in good physical condition (Images 2 and 3), are chronologically continuous, and can be retrieved at acceptable rate</td>
<td>In our department, 1936 for slides and 1956 for blocks</td>
</tr>
</tbody>
</table>

* Year of archive integrity: 1956. Optional: Evaluate the feasibility of molecular tests that will be used in potential projects.
We decided to remove the binders on the premises of our department. Next, to avoid transporting the archive for scanning, we searched for a company able to perform scanning on the premises of our department.

Evaluation of the Digital Archive

Each scanned volume was saved as a separate searchable pdf file. Overall, the digitized archive occupied 6 gigabytes. The full cost of scanning included the lease of a scanner and personnel costs for about 13 business days with an additional 2 days for OCR processing at 4 cents per page.
Pathology reports were typed using various models of typewriters, decreasing the accuracy of OCR. This problem can be compensated for by performing redundant searches, eg, to identify cases of parotid pathology, one can search for “parotid” and “salivary.” To characterize the accuracy of OCR, we searched for a term with proven occurrence: knowing that every year, starting with 1956, had at least 5,500 expected “hits,” 7 were identified. In 4 unidentified cases, “S” was recognized by the OCR software as an “S.” In the remaining 2 unidentified cases, the surgical accession number was typed over a line (rather than above the line), and “2500” was coded as “-3I!”~2.”

We defined this contextually processed code as the “digital signature.” The digital signature of any text can be easily identified by copying the text of interest from the pdf and pasting it into an html editor (Microsoft Notepad, Microsoft, Redmond, WA). Further searches can be optimized by searching for the digital signature. The following example further illustrates the concept of digital signature: In 1 case, the digital signature of “mucopidermoid” was “-ruCOEPIDER1401D.” The search for “ruCO” identified 1 additional case of MEC that was not found on initial search for “mucopidermoid.” This experience prompted us to search for shortened versions of the term of interest (eg, additional cases of MEC were identified by searching for “ucoe” and “coepi”).

The search function of the scanned archive was further tested. First, 1 volume (about 800 reports) was searched manually for salivary gland tumors. Next, the scanned version of the same volume was searched electronically, using the advanced search option in Adobe Acrobat Professional 7.0. Of 7 cases identified on manual review, all 7 were also found by electronic search. The results for any search across the entire archive (about 200 files) are available in 2 to 3 minutes when performed on a personal computer with a 2.13 GHz processor and 2 gigabytes of random-access memory.

To further characterize the quality of scanned reports, we compared the incidence of salivary MEC diagnosed in our department in 2 different periods. The first period, 1956 to 1974, is the period covered by scanned reports. The second period, 1981 to 2004, was searched by using the currently used pathology laboratory information system, and the results of that search have been reported.4 MEC was chosen because it is a relatively rare tumor and the retrieval rates for tissue blocks for salivary tumors in our department were previously characterized.5 The incidence was defined as number of MEC cases per 1,000 general surgical pathology specimens. From 1956 to 1974, we identified, on average, 1 case of MEC per 2,500 surgical specimens (57 MECs in about 138,000 general surgical pathology cases). From 1981 to 2004, 1 case of MEC was identified for every 5,400 surgical pathology cases (98 MECs in 523,000 general surgical pathology cases).

CTSI* coordinates issues of patients’ confidentiality and research. All records before 1983 were destroyed, and records after 1983 are digitized and available as electronic medical records. Privacy officer consulted about transporting records to a third party for binder removal and scanning. Binder removal and scanning were done on premises. Archived pathology records before 1970s had no social security numbers, designation of patients’ sex, dates of birth, etc; however, they should be treated as PHI. Scanned archive is not considered a research repository.

Scanned archive is stored on password-protected external hard drive with lockout after 3 incorrect password attempts.
Smith et al / Pathology Archive: Quality and Scanning

Finally, the retrieval rate for tissue blocks for rare tumors (MEC) is 75% (43 cases with available blocks of 57 cases of MEC found by searching the scanned archive). Cases with missing blocks were randomly distributed throughout the years. This rate compares favorably with retrieval rates for rare tumors in 1990 to 2005 in our department, 61%.5

Discussion

The National Bioethics Advisory Commission estimates that there are currently more than 160 million specimens in pathology archives, with about 8 million new cases added every year.6 Old pathology archives are becoming clinically irrelevant, and the financial expenses associated with their storage have to be justified. Herein, we addressed 2 issues: minimizing archive storage expenses and making an archive more attractive to researchers.

Minimizing Storage Expenses

Identification of the earliest year of archive integrity is a simple step that may lead to financial savings. One may realize immediate direct savings by discarding archive materials before the year of archive integrity. For example, the research value of reports unmatched to tissue blocks is minimal, if any. Of the 3 archive components, glass slides are the only relatively replaceable one. In theory, one could always get a recut when the block is available. However, recutting old blocks requires reembedding of the tissue (to fit modern microtomes), which is a time-consuming process requiring skilled histotechnicians and is associated with an unpleasant odor of old melted paraffin.

Paper records are an expensive way of storing information and are inconvenient for information retrieval. Some institutions have begun integrating the scanning of documents into their workflow.7 Scanning paper records for their historic or research value may further help to save on storage cost.

We highlight 2 more archive characteristics. Determining tissue block retrieval rates, as shown herein and in a previous study from our department,3 would allow a better estimate of the potential yield for research projects. Finally, if archive review is performed with a particular project in mind, one should assess the feasibility of the molecular biology technique that is essential in future projects. We have proceeded with the scanning of our archive after technically successful MECT1/MAML2 FISH was performed on tissue blocks from late 1950s. Poor DNA integrity should be taken into consideration when planning experiments on archived tissue. Review of the gross description may help to identify the fixative that was used (eg, Bouin fixative or decalcification). Even if formalin was the fixative of choice, an inconsistent approach to buffering formalin is known to compromise DNA quality. In the 1940s and 1950s, buffering of formalin was inconsistent because it was achieved by adding marble chips (calcium carbonate).8,9 A study of blocks dating back to 1952 showed that RNA quality was acceptable in about 60% of samples.10

Maximizing the Research Use of Archives

The development of a searchable pathology database from a print archive was described before11: It was focused on neuropathology specimens only, included about 50,000 scanned pages, and, being performed in a different country, did not reflect ethical and legal regulations specific to the United States. The archive created by Ehsani et al12 was used to characterize the seasonal patterns of presentation in brain tumors.

Perhaps the most prominent example of a human tissue archive is the Armed Forces Institute of Pathology repository consisting of 3 million cases with 50 million paraffin blocks.13 The uniqueness of the information in the archived material was illustrated by a study of the influenza virus responsible for the 1918 pandemic, demonstrating that this virus is closely related to early swine influenza strains.14 Pathology archives are indispensable in epidemiologic studies of associations between various oncogenes and cancers. For example, by using tissue blocks from the 1940s, Shibata et al9 showed that human papillomavirus (HPV) types 16 and 18 can be detected in cervical carcinoma at approximately the same frequency as in the 1980s.15 When assessing the impact of recently introduced HPV vaccine, this information may serve as a baseline of the HPV-related disease burden. Epidemiologic details of the link between viruses and some cancers can be better understood when analyzed over time: Does viral prevalence in a particular cancer change over time and in response to environmental carcinogens? Archived human tissue is irreplaceable in its ability to answer these questions.

We present a framework for the evaluation of archive quality and successful setup of a searchable digital archive after overcoming technical challenges and complying with ethical requirements.

From the Department of Pathology, University of Pittsburgh Medical Center, Pittsburgh, PA.

Supported by the Dr E. Leon Barnes, Consult Fund.

Address reprint requests to Dr Chiosea: Dept of Pathology, Scaife Hall A616.2, University of Pittsburgh Medical Center, 200 Lothrop St, Pittsburgh, PA 15261.

Acknowledgments: We thank Kimberly Fuhrer and Frank Fusca for excellent technical support and the Clinical and Translational Science Institute, University of Pittsburgh, for guidance.

References


