Using Systemised Nomenclature of Medicine (SNOMED) codes to select digital pathology whole slide images for long-term archiving

Mahmoud Ali ,[1,2] Harriet Evans,[1] Peter Whitney,[1] Fayyaz Minhas,[3] David R J Snead [1,4]

► Additional supplemental material for this paper are available online. To view these files, please visit the journal online (http://dx.doi.org/10.1136/jcp-2022-208483).

[1]Histopathology Department, University Hospitals Coventry and Warwickshire NHS Trust, Coventry, UK
[2]Histopathology Department, Cambridge University Hospitals NHS Foundation Trust, Cambridge, UK
[3]Department of Computer Science, University of Warwick, Coventry, UK
[4]Warick Medical School, University of Warwick, Coventry, UK

Correspondence to
Dr David R J Snead, Pathology, University Hospitals Coventry and Warwickshire NHS Trust, Coventry, Coventry, UK; david.snead@uhcw.nhs.uk

FM and DRJS are joint senior authors.

Received 13 July 2022
Accepted 20 August 2022
Published Online First 15 September 2022

To cite: Ali M, Evans H, Whitney P, *et al. J Clin Pathol* 2023;**76**:349–352.

ABSTRACT

The archiving of whole slide images represents a hurdle to digital pathology implementation largely because of the amount of data generated. The retention of glass slides is currently recommended for a minimum of 10 years, but it is for individual departments to determine how digital images are archived and for how long. In a retrospective study, we examined the combination of Systemised Nomenclature of Medicine (SNOMED) codes allocated to cases reported between July 2011 and December 2015 and recalled more than 12 months after diagnosis in comparison to non-recalled cases.
Our results show that 0.2% of cases are recalled after 12 months, and SNOMED code combinations can be used to identify which cases are likely to be recalled and which are not. This approach could reduce the number of cases archived by 62% and still ensure all cases likely to be recalled remain in the archive.

INTRODUCTION

Review of previous histology is routinely done in a large number of instances, including verification of the original diagnosis, to allow assessment of disease progression and to see if the current sample represents a new condition or relapse of a previous diagnosis.[1] In the UK, the Royal College of Pathologists (RCPath) provides guidance on the length of retention of surgical pathology slides, and in adults advises at least 10 years for slides, and 8 years for digital pathology (DP) slides.[1] Record keeping also needs to be aligned with national standards.[2]

DP is being used increasingly in the UK and worldwide. Benefits include simultaneous viewing of cases, facilitating second opinion, remote access, increased teaching and research opportunities, and the use of computer algorithms to aid assessment of slides.[3–8] Further benefits exist including the digital archiving of cases, allowing rapid retrieval of prior slides without the need to locate and retrieve them from off-site storage.[9 10] Finally, digital archives do not degrade in quality as physical slides do.[9]

Despite these benefits, the volume of data created represents a storage challenge to laboratories. DP slides vastly exceed the size of radiology image files.[9 11 12] Digital archives must be stored in a secure way, with fast access when required.[5 11]

Therefore, for departments that are currently implementing DP workflows, there are important considerations regarding the storage of digital slides. RCPath recommends retaining the glass slide as the primary reference, and states that pathology departments should determine an ‘appropriate retention policy for the digital images’, recommending retention for two laboratory inspection cycles.[13]

The additional cost of a digital archive may be inappropriate if the glass slides are to be retained. Conversely, a digital archive provides easier faster retrieval, resistance to degradation, ability to see previous case annotations and ease of sharing with colleagues.[5 7] To date, no studies have been conducted examining how the nature and content of cases may be used to focus archiving on those cases which are most likely to be reviewed again in the future. In this study, we examined Systemised Nomenclature of Medicine (SNOMED) codes from a retrospective record of slides retrieved from archive to establish of these data could be used as a basis for selecting cases for archiving.

METHOD

Case recall data

Pathology records at the University Hospital Coventry & Warwickshire NHS Trust (UHCW), which are coded at diagnosis using SNOMED V.3.5 were examined. The records of recalling slides from the offsite storage between July 2011 (when archiving offsite after 12 months started) and December 2015 inclusive when DP reporting started (43 months) were examined for specimen type, final diagnosis and the SNOMED T (Topography) and SNOMED M (Morphology) codes. This time window was chosen because the process of digitisation has reduced the need to recall the cases from the archive. Cases that had been recalled for research purposes as opposed to clinical purposes were excluded. Where recalled cases were tagged with more than one SNOMED T and M combination, only the SNOMED codes deemed most likely to have triggered the recall process were considered. This was determined by review of the clinical data.

Probability of recall

We modelled the process of case recall from archive using the SNOMED M and T codes of historical recall data. The posterior probability of recall for a case with a given M and T code $p\left(recall|\left(M,T\right)\right)$ is calculated through the Bayes rule as follows:

$$p\left(recall \mid (M,T)\right) = \frac{p\left((M,T) \mid recall\right) p\left(recall\right)}{p\left((M,T)\right)}$$

$$= \frac{\frac{N_{recall}(M,T)}{N_{recall}} \frac{N_{recall}}{N}}{\frac{N(M,T)}{N}}$$

$$= \frac{N_{recall}(M,T)}{N(M,T)}.$$

Here, $p\left((M,T) \mid recall\right)$ is the likelihood of observing the SNOMED M and T code combination in historical data of archived cases that were recalled, $p\left((M,T)\right)$ is the background probability of observing that SNOMED code combination and $p\left(recall\right)$ is the prior probability of recall irrespective of the SNOMED code. The likelihood $p\left((M,T) \mid recall\right)$ is calculated using historical data of the number of recalled cases $N_{recall}(M,T)$ with a certain SNOMED M and T code combination and the total number N_{recall} of recalled cases irrespective of SNOMED codes. The 'evidence' probability $p\left((M,T)\right)$ is taken as the ratio of $N_{recall}(M,T)$ to the total number N of cases in the archive. This allows us to express the posterior recall probability of a certain SNOMED M and T code combination as the ratio of the number of recalled cases with a certain SNOMED combination to the total number of cases observed with that combination in the archive. In order to assert our belief that cases with rare or less-frequent SNOMED combinations may be recalled with a disproportionately higher rate, the model allows addition of pseudo counts ϵ in the calculation of the 'smoothed' posterior recall probability. Since background counts and recall counts are obtained from slightly different distributions in terms of years, the denominator term in the recall probability formula is updated as follows to keep posterior probability values in the range $[0,1]$:

$$p\left(recall \mid (M,T)\right) = \frac{N_{recall}(M,T) + \epsilon}{max\left(N_{recall}(M,T), N(M,T)\right) + \epsilon}.$$

The recall probability calculated is between 0 and 1. A result of 0 means that none of the reported cases with that SNOMED code were recalled, whereas a score of 1 means that all reported cases with that SNOMED code were recalled. This probability can be used as a retention preference in an archiving solution in which cases with high expected recall probability are preferentially retained in the archive. In other words, if cases are to be deleted from the archive due to storage limitations, cases with lower retention preference will be deleted first. As a baseline, we use a naïve 'randomised storage' strategy in which all cases are equally likely to be retained in the archive irrespective of their SNOMED codes.

RESULTS

Between July 2011 and December 2015, the UHCW pathology department reported 162 761 cases and recalled 390 cases (0.2%) from the off-site store, which equates to 1 case recalled for every 417 cases reported. The recall probability of each SNOMED code combination used ranges from 0 to 1, a complete list of the SNOMED codes used and their recall rates online supplemental file 1.

Figure 1 shows the heat map of $\log_{10} \mathrm{p}\left(\mathrm{recall} \mid (\mathrm{M},\mathrm{T})\right)$ for combinations of SNOMED M and T codes with high probability of recall.

Figure 2 plots the expected recall success rate versus fraction of cases stored (retention rate) for different archiving strategies (randomised storage baseline, using recall probabilities based on M-code only, or using both M and T codes). Similar to the receiver operating characteristic (ROC) curve, which is used for measuring predictive quality of predictive models, this 'Recall-versus-Retention' (RR) curve depicts the efficiency of a certain archival strategy. It does so by showing the percentage of cases that can be expected to be recalled successfully if a certain percentage of cases is stored in the archive based on the retention preference assigned by the archival strategy. An ideal archiving strategy would store the minimum number of cases (retention rate) to give the highest expected recall success rate. As a baseline, we consider randomised archiving, in which every

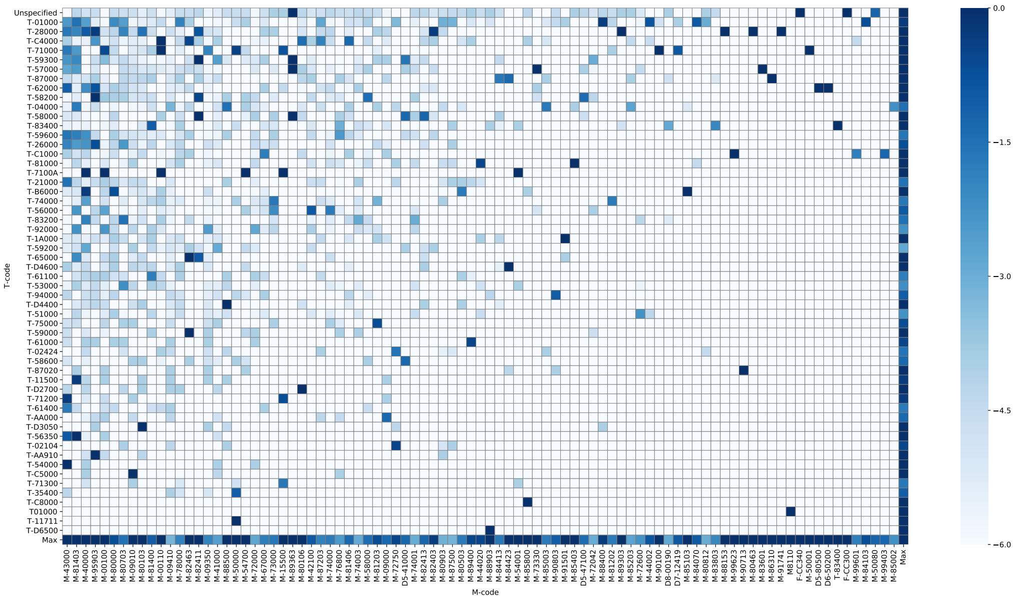

Figure 1 Heat map showing the log probability of recall for Systemised Nomenclature of Medicine (SNOMED) T and M code combinations. Blue with the blue shade proportional to the recall rate.

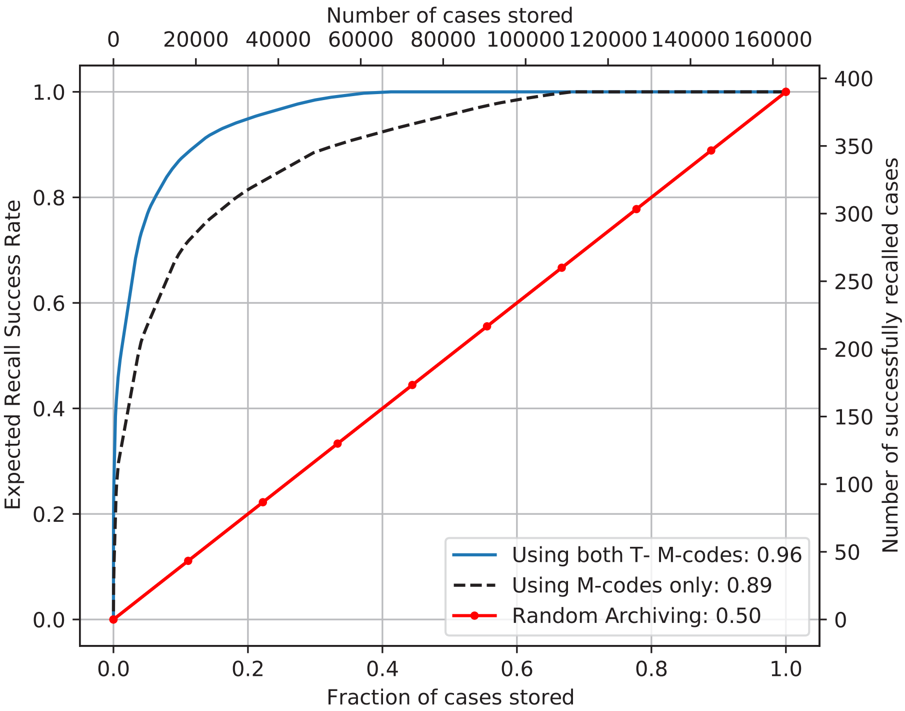

Figure 2 Recall versus retention (RR) curves for different archiving strategies showing expected case recall success rate versus fraction of cases stored based on retention preferences determined by each strategy. Numbers in the legend show the areas under the RR curves for different strategies.

case has the same recall probability and hence the same retention preference irrespective of its SNOMED code. This naïve strategy requires storing all cases in the archive to ensure a 100% recall success rate. In contrast, SNOMED code-based strategies perform better. If only SNOMED M-code data are used to derive recall probabilities, a recall success rate of 100% can be achieved by storing only 70% of cases with AURRC=0.89. Using the combination of M and T codes allows us to improve this efficiency with storage of only 38% of cases needed to deliver a 100% successful recall rate (AURRC=0.96). This shows that SNOMED codes can be effectively used as part of archiving policies in DP archival solutions.

DISCUSSION

The move to DP has resulted in different approaches to the archiving whole slide images (WSIs),[14] including keeping all data,[15] retention for 3 years[16] and no retention.[17] Other strategies proposed include removal of ×40 layer from image files[5] and using ×20 scanning of single images.[18] This variation may reflect the differences between centres desire to access data for academic purposes[5] as opposed to a purely clinical care focus.[17]

The results of this show only 0.2% of cases are recalled for clinical review after 12 months from diagnosis and that the SNOMED codes indicate which these cases are, and equally which cases are never likely to be recalled.

These results largely align with clinical expectations, although the inclusion of some malignant diagnoses such as small cell carcinoma of the lung in the never-recalled group was unexpected and indicate cases with poor outcome are unlikely to be recalled but which may nevertheless be worthy of archiving for alternative uses such as teaching and research.

This approach shows that a coding script designed to search for matches to SNOMED codes known to be recalled provides a viable approach to automate the retention of cases for archiving.

This is the first study to analyse how the SNOMED code data from prior recall of cases could be used to select WSI for archiving. With appropriate updating of code, this approach would be equally applicable to SNOMED Clinical Terms or indeed other versions of SNOMED as required. Clinical practice clearly differs between sites, so although some of the data presented may be transferable, this should be checked by some analysis of local data. The data presented provide a benchmark that could be supported by audits to validate against local practice. Such an approach would be in line with ISO 15189 standards and RCPath guidance.[1 19]

CONCLUSION

In conclusion, our study shows that SNOMED T and M codes provide a mechanism for predicting the recall probability of pathology cases from archives. Using this to select cases for archiving could help reduce the size and cost of the archive, while maintaining the advantages of easy rapid retrieval of WSI.

Handling editor Runjan Chetty.

Contributors DRJS devised the project and the main conceptual ideas. MA and FM were convinced by the presented idea. MA collected the data with the help of PW. FM worked out the technical details and performed the numerical calculations. DRJS, FM, HE and MA wrote the manuscript. DRJS supervised the project and was in charge of overall direction and planning. All authors provided critical feedback and helped shape the research, analysis and manuscript.

Funding This study is supported by the PathLAKE Centre of Excellence for digital pathology and artificial intelligence, which is funded from the Data to Early Diagnosis and Precision Medicine strand of the HM Government's Industrial Strategy Challenge Fund, managed and delivered by Innovate UK on behalf of UK Research and

Innovation. David Snead, Mahmoud Ali, and Fayyaz Minhas are part of the PathLAKE consortium and all participated in this work. David Snead devised the project and the main conceptual ideas. Mahmoud Ali collected the data. Fayyaz Minhas worked out the technical details and performed the numerical calculations. David Snead, Fayyaz Minhas, and Mahmoud Ali participated in writing the manuscript.

Competing interests DRJS is co-owner, director and shareholder of Histofy AI. All the other contributors do not have conflicts of interest to declare.

Patient consent for publication Not applicable.

Ethics approval This study was conducted under National Research Ethics Service approval 15/NW/0843; IRAS 189095

Provenance and peer review Not commissioned; internally peer reviewed.

Data used in this study is available by application to PathLAKE https://www.pathlake.org

ORCID iDs

Mahmoud Ali http://orcid.org/0000-0001-7722-6196
David R J Snead http://orcid.org/0000-0002-0766-9650

REFERENCES

1 Wilkins B. *The retention and storage of pathological records and specimens*. 5th edition. Royal College of Pathologists: Royal College of Pathologists, 2015.
2 NHSX. Records managment code of practice. In: London, ed. *NHSX*, 2021.
3 Snead DRJ, Tsang Y-W, Meskiri A, *et al*. Validation of digital pathology imaging for primary histopathological diagnosis. *Histopathology* 2016;68:1063–72.
4 Williams BJ, Brettle D, Aslam M, *et al*. Guidance for remote reporting of digital pathology slides during periods of exceptional service pressure: an emergency response from the UK Royal College of pathologists. *J Pathol Inform* 2020;11:12.
5 Stathonikos N, Nguyen TQ, Spoto CP, *et al*. Being fully digital: perspective of a Dutch academic pathology laboratory. *Histopathology* 2019;75:621–35.
6 Mukhopadhyay S, Feldman MD, Abels E, *et al*. Whole slide imaging versus microscopy for primary diagnosis in surgical pathology: a multicenter blinded randomized Noninferiority study of 1992 cases (pivotal study). *Am J Surg Pathol* 2018;42:39–52.
7 Pallua JD, Brunner A, Zelger B, *et al*. The future of pathology is digital. *Pathol Res Pract* 2020;216:153040.
8 Salto-Tellez M, Maxwell P, Hamilton P. Artificial intelligence-the third revolution in pathology. *Histopathology* 2019;74:372–6.
9 Huisman A, Looijen A, van den Brink SM, *et al*. Creation of a fully digital pathology slide archive by high-volume tissue slide scanning. *Hum Pathol* 2010;41:751–7.
10 Hanna MG, Reuter VE, Samboy J, *et al*. Implementation of digital pathology offers clinical and operational increase in efficiency and cost savings. *Arch Pathol Lab Med* 2019;143:1545–55.
11 Pantanowitz L, Sharma A, Carter AB, *et al*. Twenty years of digital pathology: an overview of the road travelled, what is on the horizon, and the emergence of Vendor-Neutral Archives. *J Pathol Inform* 2018;9:40.
12 Turnquist C, Roberts-Gant S, Hemsworth H, *et al*. On the edge of a digital pathology transformation: views from a cellular pathology laboratory focus group. *J Pathol Inform* 2019;10:37.
13 Cross S, Furness P, Igali L. *Best practice recomendations for implementing digtial pathology*. Royal college of pathologists: RCPath, 2018.
14 Thorstenson S, Molin J, Lundström C. Implementation of large-scale routine diagnostics using whole slide imaging in Sweden: digital pathology experiences 2006-2013. *J Pathol Inform* 2014;5:14.
15 Stathonikos N, Nguyen TQ, van Diest PJ. Rocky road to digital diagnostics: implementation issues and exhilarating experiences. *J Clin Pathol* 2021;74:415–20.
16 Retamero JA, Aneiros-Fernandez J, Del Moral RG. Complete digital pathology for routine histopathology diagnosis in a multicenter Hospital network. *Arch Pathol Lab Med* 2020;144:221–8.
17 Baidoshvili A. How to go digital in pathology, 2016. LabPON Laboratorium Pathologie Oost-Nederland. Available: https://www.philips.com/c-dam/b2bhc/master/sites/pathology/resources/white-papers/labron-how-to-go-digital.pdf
18 Eloy C, Vale J, Curado M, *et al*. Digital pathology workflow implementation at IPATIMUP. *Diagnostics* 2021;11. doi:10.3390/diagnostics11112111. [Epub ahead of print: 15 11 2021].
19 International Organization for Standardization. Iso 15189:2012 medical laboratories — requirements for quality and competence 2012.