OPTIMAM Mammography Image Database: A Large-Scale Resource of Mammography Images and Clinical Data

Mark D. Halling-Brown, BSc, MSc, PhD • Lucy M. Warren, BSc, MSc, PhD • Dominic Ward, BSc, PhD • Emma Lewis, BSc, MSc, PhD • Alistair Mackenzie, BSc, MSc, PhD • Matthew G. Wallis, MB, ChB, FRCR • Louise S. Wilkinson, BA, BM, BCh, FRCR • Rosalind M. Given-Wilson, MBBS, MRCP, FRCR • Rita McAvinchey, BM, MBBS, MSc • Kenneth C. Young, BSc, PhD

From the Department of Scientific Computing (M.D.H.B., D.W., E.L.) and National Co-ordinating Centre for the Physics of Mammography (L.M.W., A.M., K.C.Y.), Royal Surrey NHS Foundation Trust, Egerton Road, Guildford GU2 7XX, England; Centre for Vision, Speech and Signal Processing (M.D.H.B., E.L.) and Department of Physics (K.C.Y.), University of Surrey, Guildford, England; Cambridge Breast Unit, Cambridge University Hospitals NHS Foundation Trust, Cambridge, England (M.G.W.); NIHR Cambridge Biomedical Research Centre, Cambridge, England (M.G.W.); Oxford Breast Imaging Centre, Oxford University Hospitals NHS Foundation Trust, Oxford, England (L.S.W.); Department of Radiology, St George's Healthcare NHS Trust, London, England (R.M.G.W.); and Jarvis Breast Screening Centre, Guildford, England (R.M.). Received May 18, 2020; revision requested July 7; revision received September 3; accepted October 5. **Address correspondence to** M.D.H.B. (e-mail: *mhalling-brown@nhs.net*).

The creation and ongoing development of the OPTIMAM Image Database is funded by Cancer Research UK (C30682/A28396).

Conflicts of interest are listed at the end of this article.

Supplemental material is available for this article.

Radiology: Artificial Intelligence 2021; 3(1):e200103 • https://doi.org/10.1148/ryai.2020200103 • Content codes: AI BR • ©RSNA, 2021

The development of artificial intelligence software to improve the outcomes of breast screening relies on the availability of well-curated image databases (1). The OPTIMAM Mammography Image Database (OMI-DB) (2,3) was created to provide a centralized, fully annotated dataset for research. The initial reason for creating the database was for the Cancer Research United Kingdom–funded projects OPTIMAM (2008–2013) and OPTIMAM2 (2013–2018), which evaluated how various factors affect breast cancer detection on mammograms. The images are derived from screening centers in the United Kingdom and combined with systematically collected data on the current screening episode, as well as previous and subsequent episodes. In the United Kingdom, the National Health Service Breast Screening Programme (NHSBSP) invites women to attend breast screening every 3 years between the ages of 50 and 70 years. A screening episode is one attendance at screening by a woman and includes any immediate workup imaging (assessment) if she was recalled for further investigation of a suspicious region on the screening mammograms. Any pathologic finding is also included, and the episode ends with histologic diagnosis or treatment for all lesions. At some screening centers younger and older women are also invited for screening as part of the national age trial (4). Some women in high-risk groups receive annual invitations to screening. Our objective was to collect mammograms for women with screen-detected cancers as well as representative samples of normal and benign screening cases.

Materials and Methods

Content of OMI-DB

"For processing" and "for presentation" screening mammograms and prior mammograms have been collected for all screen-detected and interval cancers from Jarvis Breast Screening Centre in Guildford, St George's Hospital in South West London, and Addenbrooke's Hospital in Cambridge since 2011. All mammography images and data associated with initial screening attendance, further assessment, and surgical outcomes were collected as a screening episode. In addition to continuous collection of cancers, images and clinical data were collected for all women screened during 2014, and for a random selection of 25% of all women screened in 2012, 2013, and 2015 at two of the three sites. The total number of all types of images in the database is 3 072 878. Collection into the database is ongoing, and each case is updated with new information and further screening episodes. All data in this Data Resource article relate to May 2020. Data from a total of 172 282 women were included within the database at this time.

Image Database: Image Collection and Design

The processes and systems required for image collection are complex. A key provision is to ensure that all potentially identifiable information is removed from the images and data at the point of collection and is inaccessible to researchers. Figure 1 shows a simplified view of two types of collection: automated remote site and stand alone. A full description can be found in Appendix E1 (supplement). To identify women's data for collection, the National Breast Screening System (NBSS) is queried using search criteria such as study-date range or outcome classification (normal or malignant). Images and clinical data for the women's screening and assessment episodes are then retrieved. For the images currently in OMI-DB, this process is fully automated using a physical or virtual server. In addition, tools have been developed and tested for small-scale collection sites where setting up a server is problematic. In this stand-alone system, an image collection tool is downloaded by a staff member and pointed at a manually prepared folder containing images for collection, and the NBSS queried for clinical data for those women. Further processing and storage of images and data follows the automated collection procedure. Imag-

Abbreviations

NBSS = National Breast Screening System, NHSBSP = National Health Service Breast Screening Programme, OMI-DB = OPTIMAM Mammography Image Database

Summary

The OPTIMAM Mammography Image Database is a sharable resource with processed and unprocessed mammography images from United Kingdom breast screening centers, with annotated cancers and clinical details.

Key Points

- The database includes serial screening mammograms that were collected over a 10-year period with data from 172 282 women as of May 2020.
- The database includes data on all breast cancers in a screened population including interval cancers.
- This resource has been widely used to develop and evaluate artificial intelligence algorithms for breast cancer detection.

ing and screening data are pseudonymized and records inserted into lookup tables at the collection site. Images, metadata, and screening data are uploaded to the cloud for storage. Activity relating to the collection of data and images is logged at the collection site.

The OMI-DB comprises several relational databases and cloud storage systems (3). Figure 2 shows a simplified schema of the data model. The associated data comprise radiologic, clinical, and pathologic information extracted from NBSS. When loading images into the database, relevant Digital Imaging and Communications in Medicine tags are extracted to create a searchable index. Information on screening history, previous occurrences of cancer, biopsy results, and surgical procedures are collected from NBSS. The exact radiologic locations of lesions are not stored in NBSS. However, such information, important for training and evaluating algorithms, is collected in OMI-DB. Experienced (UK accredited) mammography readers at their own site (radiologists and advanced practice radiographers) annotate the images with reference to records made at the time of initial mammography interpretation and at further (assessment) workup (magnification views, US, and biopsy). This information is used to define rectangular regions of interest indicating the location and area of lesions and other attributes, such as radiologic appearance and conspicuity. To be accredited, readers must have screening and workup experience reading a minimum of 5000 cases per year. These local annotations are preferred to external review without access to the local information. Of the 7661 screen-detected cancers from episodes with digital images, 5097 have been marked (67% of cancers). Web-enabled, remotely accessible software allows radiologists to view cases, annotate clinical features, and participate in observer studies (5).

Data Sharing

The project has approval from an ethical research committee specializing in research databases organized by the NHS Health Research Authority. A formal local agreement to contribute cases is also gained at each site. The funder, Cancer Research UK, retains the intellectual property of the database and implements sharing agreements with approved academic and commercial research groups. The sharing of images and metadata with external parties involves additional processes such as regeneration of pseudonyms and recording cases that have been shared with each third party. A download-coordinating tool facilitates secure transfer and regularly synchronizes the metadata defined by the access list. The tool is supplied with documentation providing an overview of the database structure and tables describing the relational schema. Researchers are encouraged to use our open-source Python package that parses the shared OMI-DB to provide a

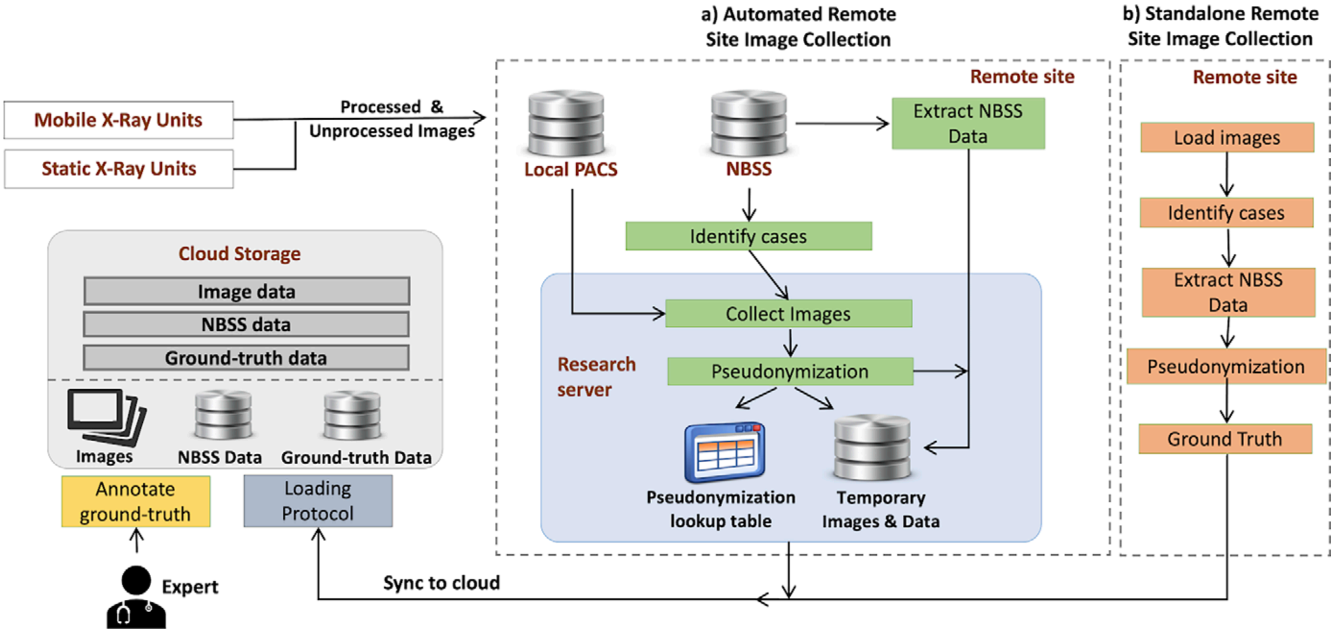

Figure 1: Simplified representation of processes for collecting and annotating imaging, clinical, and ground truth data used to populate the OPTIMAM Mammography Image Database. NBSS = National Breast Screening System, PACS = picture archiving and communication system.

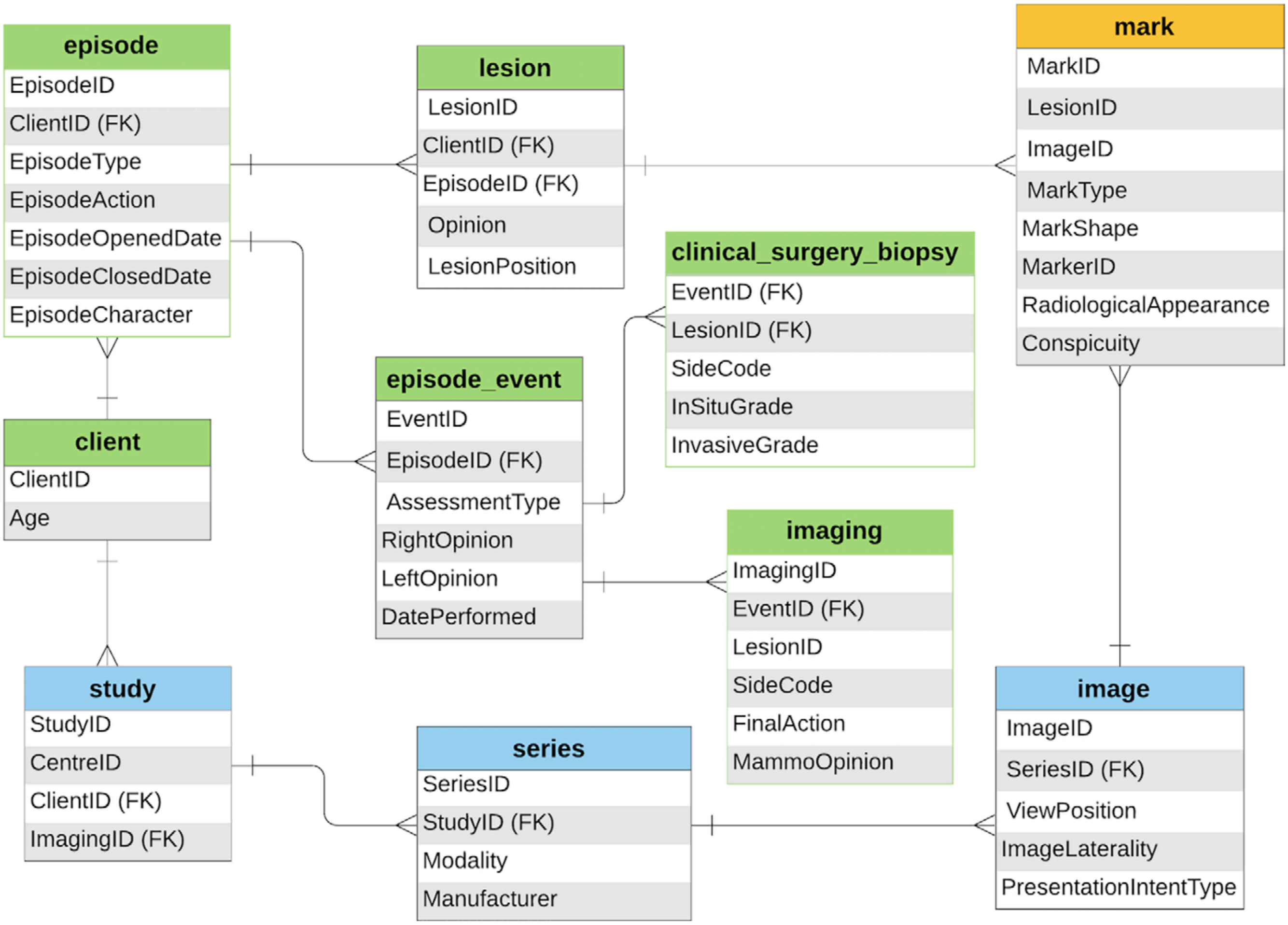

Figure 2: Schema shows simplified data model for radiologic (blue), clinical and pathologic (green), and ground truth (orange) information stored in the OPTIMAM Mammography Image Database. FK = foreign key.

Table 1: Status and Grade of Screen-detected and Interval Cancers in OMI-DB

Invasive Status/Grade	No. of Screen-detected Cancers	No. of Interval Cancers
Invasive		
Grade 1	1222	71
Grade 2	2818	367
Grade 3	977	247
Not assessable	24	7
No grade	487	177
In situ		
Low grade	166	9
Intermediate grade	589	20
High grade	931	36
Not assessable	0	0
No grade	420	18
Total	7634	952

Note.—For details of the classification procedures and source code used to generate the count data, visit *https://bitbucket.org/scicomcore/omidb-2020*. OMI-DB = OPTIMAM Mammography Image Database.

detailed application programming interface and tools to facilitate metadata extraction and filtering. For access to the data and images from OMI-DB, contact the Cancer Research UK's Commercial Partnerships team or apply on the OMI-DB website (6).

Resulting Dataset and Applications

Patient Characteristics and Database Overview

Table 1 shows the invasive status and grade for screen-detected and interval cancers in general population screening, for which OMI-DB contains the associated digital images. Twenty-seven screen-detected cancers and 14 interval cancers were excluded from the tables due to missing, or inconsistent, information. The number of women with one, two, or three or more screening episodes with digital images in OMI-DB is shown in Figure 3. Note that OMI-DB also contains clinical information for earlier analog imaging episodes but no images.

OMI-DB contains screening images for 172 282 women attending breast screening (including 4518 women who also have tomosynthesis assessment images) and screening images for 364 women with higher-risk

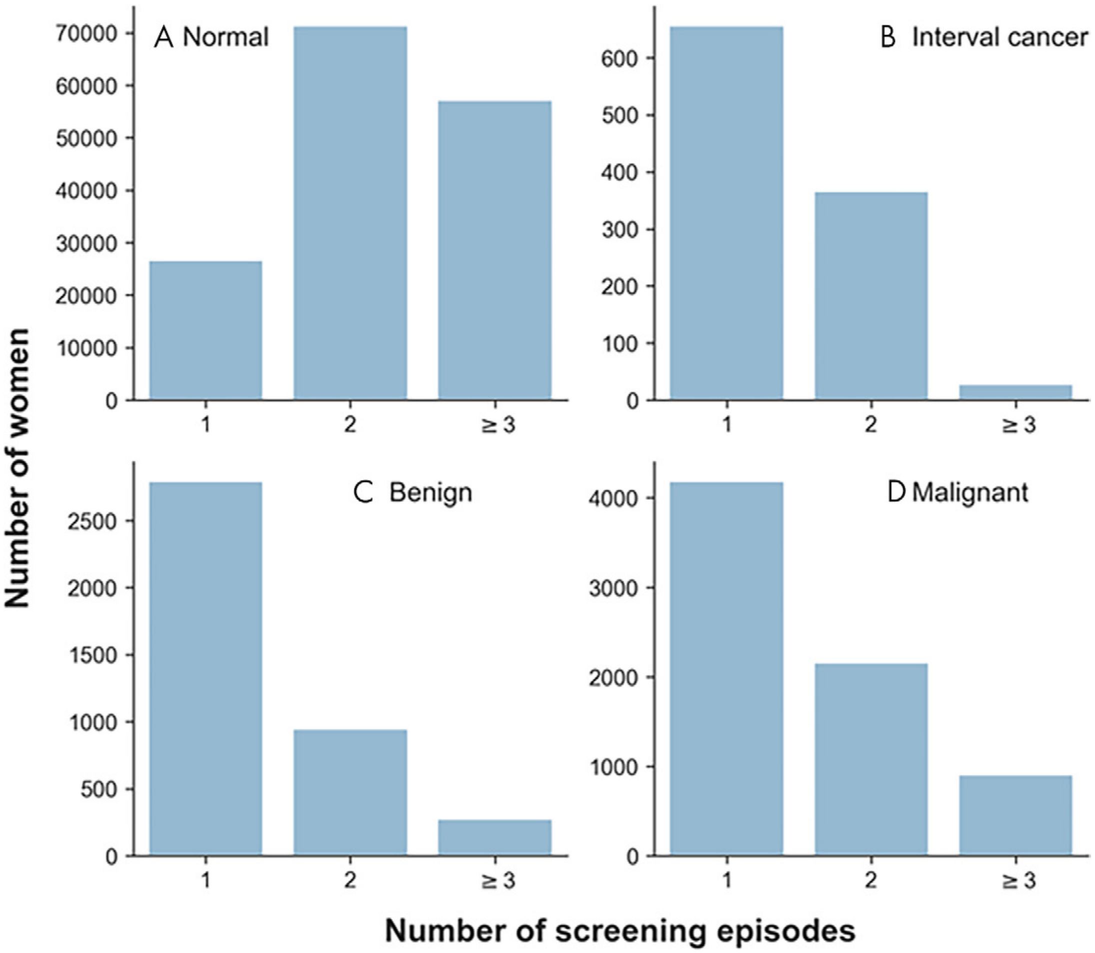

Figure 3: Number of women with one, two, or three or more screening episodes with images in OPTIMAM Mammography Image Database, for, *A*, women with normal breasts, *B*, women with interval cancers, *C*, women with benign lesions, and, *D*, women with screen-detected cancers. For cases of nonnormal breasts, episodes beyond the most recent abnormal episode were excluded.

screening episodes. The number of episodes in OMI-DB for each age group is shown in Table 2. As the collection has progressed, modalities, such as tomosynthesis and MRI, have been added.

Use of Database

The database has been used in projects at the Royal Surrey NHS Foundation Trust (7–10). This includes virtual clinical trials investigating the effect of factors such as detector type, dose, and image processing on breast cancer detection (9,10) and evaluating the cancer characteristics and breast density of women in the NHSBSP (8,11).

The database has been shared since 2014, and publications by users must acknowledge its use. Cancer Research UK staff require users to report on their project's progress even if not published. Data and images from OMI-DB have been shared with more than 30 academic and commercial groups for various research aims, but mainly to develop machine learning artificial intelligence techniques (12–16). Images and data have also been used to evaluate several artificial intelligence algorithms at different stages of development from prototypes to commercial products (12–15).

Table 2: Number of Episodes for Routine Screening and for the Higher Risk Group in OMI-DB

Age Group (y)	Routine Screening by Age	Higher Risk
30–34	0	3
35–39	0	8
40–44	2	226
45–49	19 545	219
50–54	90 699	193
55–59	88 923	126
60–64	80 067	83
65–69	72 774	74
70–74	21 068	37
75–79	18	11
80–84	1	3
Total	373 097	983

Discussion

The set-up of an annotated mammographic image database with sharing protocols has been time-consuming and challenging. Any collection process should ideally be automatic, link clinical data to images, while retaining confidentiality and expert annotation. Since 2011, we have met these challenges and created a large image database of the full range of cases acquired during breast screening. Live updates on the data in Table 1 and Figure 3 are provided at the database website (6). The database has sharing protocols that allow images to be used by researchers around the world (12–16).

The availability of sequential screening events and interval cancers opens up many artificial intelligence research applications, including whether an abnormality could have been detected earlier. Sequential normal screening mammograms can be analyzed using quantitative imaging features, with a priori knowledge that some years later particular cases develop a malignancy.

A limitation of the dataset is that UK screening is based on invitations to women aged 50 to 70 years, and few younger women are therefore included. Since 2010 as part of the AgeX trial (17), first invitations occur in the age range 45–52 years. Women older than 70 years can self-refer.

Overall, a valuable, sharable database has been developed which holds both processed and unprocessed mammography images with annotated cancers and clinical details.

Acknowledgments: The authors would like to thank the Jarvis Breast Screening Centre, The Rose Centre South West London Breast Screening at St George's Hospital, and Cambridge Breast Unit at Addenbooke's Hospital for allowing copies of their images to contribute to OMI-DB. The authors would also like to thank Dr Vicky Cooke, Dr Julie Cooke, Dr Anna Cummin, Charul Patel, Pat Kelly, Jacqueline Kenny, and Judy Leo for annotating the lesions within OMI-DB.

Author contributions: Guarantors of integrity of entire study, M.D.H.B., D.W., K.C.Y.; study concepts/study design or data acquisition or data analysis/interpretation, all authors; manuscript drafting or manuscript revision for important intellectual content, all authors; approval of final version of submitted manuscript, all authors; agrees to ensure any questions related to the work are appropriately resolved, all authors; literature research, M.D.H.B., E.L., A.M.; clinical studies, L.M.W., R.M.; experimental studies, M.D.H.B.; statistical analysis, M.D.H.B., L.M.W., D.W., E.L.; and manuscript editing, M.D.H.B., L.M.W., D.W., E.L., A.M., M.G.W., L.S.W., R.M.G.W., K.C.Y.

Disclosures of Conflicts of Interest: **M.D.H.B.** Activities related to the present article: The development of the database and its ongoing management was and is funded by grants from Cancer Research UK to The Royal Surrey NHS Foundation Trust. C30682/A28396. Activities not related to the present article: Royal Surrey NHS Foundation Trust received funding from Google Health UK to support research collaboration in the development and evaluation of Google's AI system for breast cancer screening (McKinney et al. International evaluation of an AI system for breast cancer screening. Nature. Nature Publishing Group; 2020;577(7788):89–94. doi:10.1038/s41586-019-1799-6); Royal Surrey NHS Foundation Trust receives additional income from Cancer Research UK as a share of payments for commercial licensed access to the OPTIMAM Mammography Image Database. Other relationships: disclosed no relevant relationships. **L.M.W.** Activities related to the present article: The development of the database and its ongoing management was and is funded by grants from Cancer Research UK to The Royal Surrey NHS Foundation Trust. Activities not related to the present article: Royal Surrey NHS Foundation Trust received funding from Google Health UK to support research collaboration in the development and evaluation of Google's AI system for breast cancer screening (McKinney et al. International evaluation of an AI system for breast cancer screening. Nature. Nature Publishing Group; 2020;577(7788):89–94. doi:10.1038/s41586-019-1799-6); author teaches two lectures a year on a MSc course at the University of Surrey; Royal Surrey NHS Foundation Trust receives additional income from Cancer Research UK as a share of payments for commercial licensed access to the OPTIMAM Mammography Image Database. Other relationships: disclosed no relevant relationships. **D.W.** Activities related to the present article: The development of the database and its ongoing management was and is funded by grants from Cancer Research UK to The Royal Surrey NHS Foundation Trust. Activities not related to the present article: Royal Surrey NHS Foundation Trust received funding from Google Health UK to support research collaboration in the development and evaluation of Google's AI system for breast cancer screening (McKinney et al. International evaluation of an AI system for breast cancer screening. Nature. Nature Publishing Group; 2020;577(7788):89–94. doi:10.1038/s41586-019-1799-6); Royal Surrey NHS Foundation Trust receives additional income from Cancer Research UK as a share of payments for commercial licensed access to the OPTIMAM Mammography Image Database. Other relationships: disclosed no relevant relationships. **E.L.** Activities related to the present article: The development of the database and its ongoing management was and is funded by grants from Cancer Research UK to The Royal Surrey NHS Foundation Trust. Activities not related to the present article: Royal Surrey NHS Foundation Trust received funding from Google Health UK to support research collaboration in the development and evaluation of Google's AI system for breast cancer screening (McKinney et al. International evaluation of an AI system for breast cancer screening. Nature. Nature Publishing Group; 2020;577(7788):89–94. doi:10.1038/s41586-019-1799-6); Royal Surrey NHS Foundation Trust receives additional income from Cancer Research UK as a share of payments for commercial licensed access to the OPTIMAM Mammography Image Database. Other relationships: disclosed no relevant relationships. **A.M.** disclosed no relevant relationships. **M.G.W.** Activities related to the present article: Cancer Research UK program grant partly funded setup of database. Activities not related to the present article: disclosed no relevant relationships. Other relationships: disclosed no relevant relationships. **L.S.W.** disclosed no relevant relationships. **R.G.W.** Activities related to the present article: Cancer Research UK charity grant funding the OPTIMAM project staff undertaking data collection from 2009 to 2019. Activities not related to the present article: disclosed no relevant relationships. Other relationships: disclosed no relevant relationships. **R.M.** disclosed no relevant relationships. **K.C.Y.** Activities related to the present article: The development of the database and its ongoing management was and is funded by grants from Cancer Research UK to The Royal Surrey NHS Foundation Trust. Activities not related to the present article: Royal Surrey NHS Foundation Trust received funding from Google Health UK to support research collaboration in the development and evaluation of Google's AI system for breast cancer screening (McKinney et al. International evaluation of an AI system for breast cancer screening. Nature. Nature Publishing Group; 2020;577(7788):89–94. doi:10.1038/s41586-019-1799-6); Royal Surrey NHS Foundation Trust receives additional income from Cancer Research UK as a share of payments for commercial licensed access to the OPTIMAM Mammography Image Database. Other relationships: disclosed no relevant relationships.

References

1. Debelee TG, Schwenker F, Ibenthal A, Yohannes D. Survey of deep learning in breast cancer image analysis. Evol Syst 2020;11:143–163.
2. Patel MN, Looney P, Young KC, Halling-Brown MD. Automated collection of medical images for research from heterogeneous systems: trials and tribulations. In: Law MY, Cook TS, eds. Proceedings of SPIE: medical imaging 2014—PACS and imaging informatics: next generation and innovations. Vol 9039. Bellingham, Wash: International Society for Optics and Photonics, 2014; 90390C.
3. Halling-Brown MD, Looney PT, Patel MN, Mackenzie A, Young KC. The oncology medical image database (OMI-DB). In: Law MY, Cook TS, eds. Proceedings of SPIE: medical imaging 2014—PACS and imaging informatics: next generation and innovations. Vol 9039. Bellingham, Wash: International Society for Optics and Photonics, 2014; 903906.
4. Patnick J. Nationwide cluster-randomised trial of extending the NHS breast screening age range in England. https://doi.org/10.1186/ISRCTN33292440. Published 2016. Accessed April 8, 2020.
5. Looney PT, Young KC, Halling-Brown MD. MedxViewer: providing a web-enabled workstation environment for collaborative and remote medical imaging viewing, perception studies and reader training. Radiat Prot Dosimetry 2016;169(1-4):32–37.
6. OMI-DB OPTIMAM Mammography Imaging. https://medphys.royalsurrey.nhs.uk/omidb/. Accessed September 3, 2020.
7. Patel MN, Looney PT, Young KC, Halling-Brown MD. Quantitative imaging features: extension of the oncology medical image database. In: Cook TS, Zhang J, eds. Proceedings of SPIE: medical imaging 2015—PACS and imaging informatics: next generation and innovations. Vol 9418. Bellingham, Wash: International Society for Optics and Photonics, 2015; 941812.
8. Warren LM, Halling-Brown MD, Wilkinson LS, et al. Changes in breast density. In: Nishikawa RM, Samuelson FW, eds. Proceedings of SPIE: medical imaging 2019—image perception, observer performance, and technology assessment. Vol 10952. Bellingham, Wash: International Society for Optics and Photonics, 2019; 109520W.

9. Mackenzie A, Warren LM, Wallis MG, et al. Breast cancer detection rates using four different types of mammography detectors. Eur Radiol 2016;26(3):874–883.
10. Warren LM, Halling-Brown MD, Looney PT, et al. Image processing can cause some malignant soft-tissue lesions to be missed in digital mammography images. Clin Radiol 2017;72(9):799.e1–799.e8.
11. Burnside ES, Warren LM, Wilkinson LS, Young KC, Myles J, Duffy SW. Using Quantitative Breast Density Analysis to Predict Interval Cancers and Node Positive Cancers in Pursuit of Improved Screening Protocols. Radiological Society of North America. http://archive.rsna.org/2018/18010897.html. Published 2018. Accessed April 8, 2020.
12. McKinney SM, Sieniek M, Godbole V, et al. International evaluation of an AI system for breast cancer screening. Nature 2020;577(7788):89–94.
13. Halling-Brown M, Rodrigues-Ruiz A, Karssemeijer N, Wallis MG, Young KC. Artificial Intelligence Detecting Breast Cancer in a Screening Population: Accuracy, Earlier Detection on Prior Mammograms, and Relation with Cancer Grade. Oak Brook, Ill: Radiological Society of North America, 2019.
14. Kim HE, Kim HH, Han BK, et al. Changes in cancer detection and false-positive recall in mammography using artificial intelligence: a retrospective, multireader study. Lancet Digit Health 2020;2(3):E138–E148.
15. Agarwal R, Díaz O, Yap MH, Lladó X, Martí R. Deep learning for mass detection in Full Field Digital Mammograms. Comput Biol Med 2020;121:103774.
16. Alam N, Denton ERE, Zwiggelaar R. Classification of Microcalcification Clusters in Digital Mammograms Using a Stack Generalization Based Classifier. J Imaging 2019;5(9):76.
17. NHS Breast Screening Programme. AgeX Trial. http://www.agex.uk/. Accessed September 3, 2020.