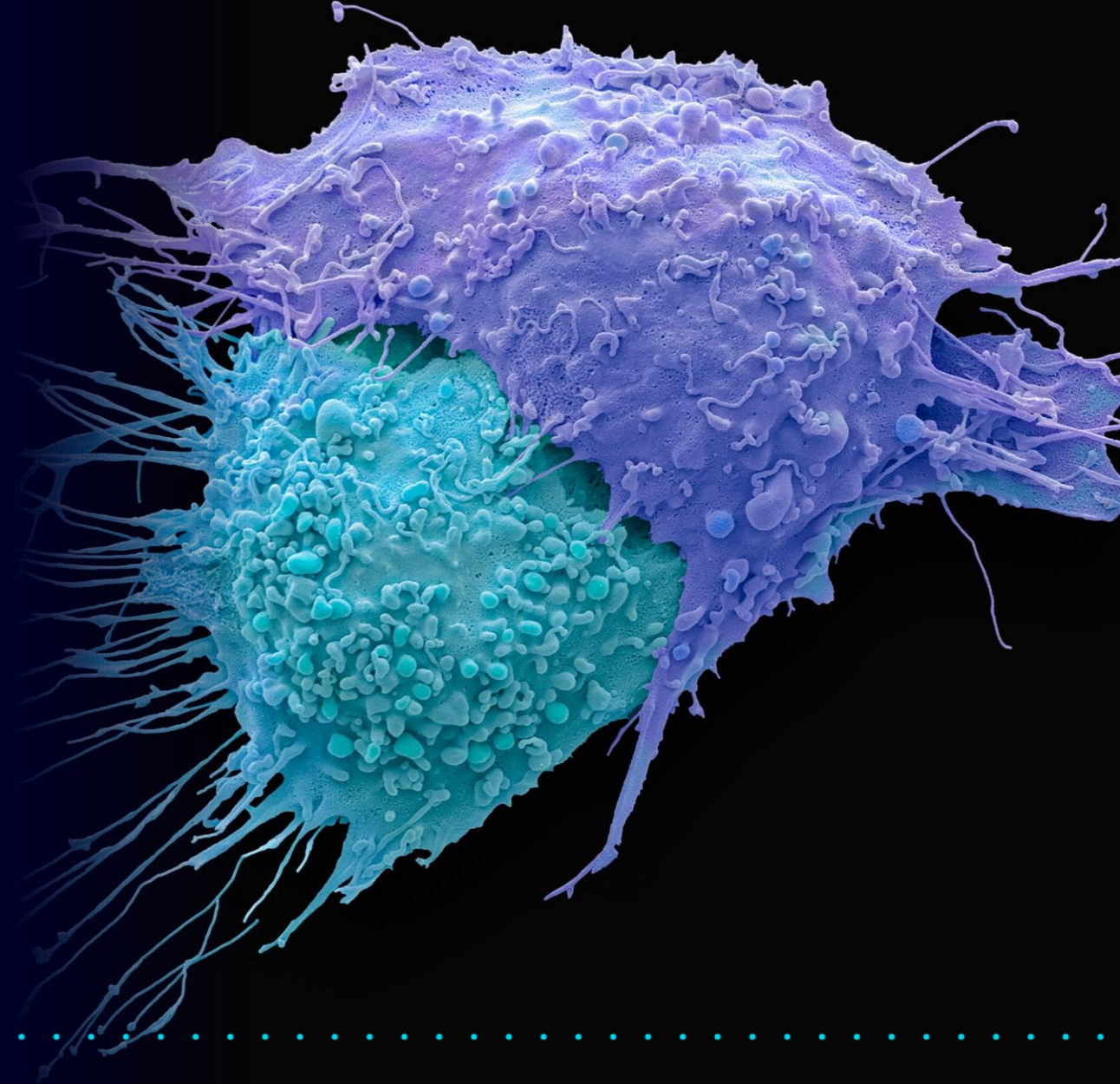# Overview

- The OPTIMAM Database (OMI-DB) is a large, centralised and de-identified database of **mammography images** collected from **multiple NHS breast screening sites** across the UK

- The ongoing collection of images and management of the database is conducted by the **Royal Surrey NHS Foundation Trust** and is funded with support from **Cancer Research UK (CRUK)**

- OMI-DB collects both processed and unprocessed **mammogram images**, **associated clinical data** and **annotations** related to screen-detected cancers and interval cancers, as well as expert-determined region of interest (ROI) annotations of malignant and benign cases

- Where available, follow-up assessment via **3D screening (digital breast tomosynthesis)** assessment and magnetic resonance imaging (MRI) of very high-risk cases is also collected

- OMI-DB has been designed to be shared for research purposes to help overcome the lack of large, curated and representative databases of medical images, which is a barrier to medical imaging research. CRUK's innovation engine, **Cancer Research Horizons** (CRH) facilitates the licencing of the OPTIMAM data

- Applications to access OPTIMAM data for research studies can be made to our **Data Access Committee** (DAC) and are open to researchers from healthcare institutions, academic centres, commercial or non-profit organisations

Database Size – 2D Mammograms

The database contains mammography images collected since 2011, with **153,418 2D** Cases available for sharing (*Aug 23*)

Type	#
Normal	150,000
Malignant (marked)*	7,096
Malignant (unmarked)	2,565
Benign (marked)*	1,400
Benign (unmarked)	5,558
Interval	1,916

Bespoke requests can be accommodated, and data requirements discussed with CRH & the OPTIMAM team, subject to Data Access Committee (DAC) approval

Malignant and Benign cases are sub-classified into marked and unmarked, depending on whether ROIs have been annotated by an expert mammography reader

Interval Cases are those that were detected in the 3-year interval between screening episodes

Database Size – 3D Tomosynthesis

The database currently contains **5,790 Tomosynthesis (3D)** screening episodes available for sharing (Aug 23):

Type	#
Normal*	2,625
Malignant (Marked)	1,568
Malignant (Unmarked)	653
Benign** (Marked)	130
Benign (Unmarked)	867
Interval	163

Digital breast tomosynthesis (DBT) are collected where available when a case is recalled for further assessment. Images are typically of a single breast or one view of the breasts where further assessment was required.

The accompanying prior 2D screening images are annotated with the ROIs and made available alongside the DBT scan

Normal cases include those recalled for assessment and recalled for assessment plus proven normal with biopsy

** *Benign cases are proven with biopsy*

Data Description

ANNOTATED IMAGES

- Expert radiologists at the screening sites retrospectively identify lesions on the mammogram images and draw a rectangular ROI around its boundaries representing ground truth (bounding box)

- Coordinates and various properties of the tumour/lesions are recorded and described in the associated image data

UNPROCESSED IMAGES

- Special arrangements have been made at image collection sites so that unprocessed images are not discarded, whereas most PACS systems used in the NHS only store processed images

- Considerable amounts of unprocessed images enable research into the fundamental physics of medical imaging

Data Description

ASSOCIATED CLINICAL DATA

- Invasive status
- Disease grade
- Mutation status
- Screen Reader opinion
- Size of tumour
- Hormone replacement status

- *For full details please see description of contents*

LONGITUDINAL DATA

- New mammography images and data are continually added to OMI-DB adding to the richness of longitudinal data available

- Information on previous screening episodes and interval cancers are available within the database and can be longitudinally linked

Data Description

TECHNICAL & BIOLOGICAL FEATURES

- Cases are recorded and classified with any technical or biological features such as microcalcifications or image blurring that can enable further investigation and research

- In addition to the clinical information, information regarding breast density may be requested (sharing is subject to further approval)

  - Breast density is calculated by Volpara Density (a fully automated breast density assessment software)

Automated Data Collection

PACS
- Processed & unprocessed images are stored on the Picture Archiving & Communication System (PACS) at each screening site

Query NBSS
- The National Breast Screening System (NBSS) clinical database at each screening site is queried to identify cases for collection

Data Extraction
- Mammogram images & the associated clinical data are automatically extracted from the site's PACS to the site's research server

Pseudony misation
- Images & clinical data are de-identified, a pseudonym is assigned & stored in a local database on the site's research server
- All identifiable images & data are permanently deleted from the site's research server

Cloud Upload
- Pseudonymised images & clinical data are uploaded to a secure, centralised cloud database, after which all the initial de-identified images & clinical data on the research site's server are deleted

Data Sharing
- Upon approval of a new research proposal, an appropriate selection of pseudonymised images & clinical data are downloaded from the cloud by the database managers for further research purposes

Equipment Manufacturer Distribution

The equipment manufacturer distribution within OMI-DB is as follows:

For **2D** screening (Jun 22):

Manufacturer	Proportion of Total Images within Database
Hologic	91.59%
GE Healthcare	4.94%
Siemens	0.72%
Philips	0.19%
Others	2.56%

For **3D** follow-up assessment (Aug 23):

| Manufacturer	Proportion of Total Images within Database
Hologic	93.85%
Siemens	6.10%
GE Medical Systems	0.05%

Expansion of data collection and further onboarding of new screening centres aims to achieve a wider diversity of X-ray equipment manufacturers. We anticipate this data to be available for sharing in the next few years.

Use Cases
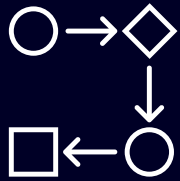
Training and Testing of Deep Learning Computer Aided Detection Algorithms

Considerably aided by the presence of a large number of well-labelled malignant cases as well as details of interval cancers & large numbers of normal cases
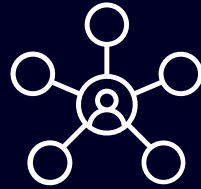
Training of Models Designed to Predict Risk & Stratify Patient Populations

Supported by the existence of sequential screening episodes and known outcomes over a long time period
(9+ years)
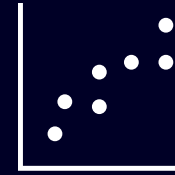
Density Investigation & Fundamental Physics of Image formation

Technical repeats and calculated breast density make investigations into density & image quality possible, alongside unprocessed data to explore fundamental physics of imaging

Pre-Market Validation Exercises

Companies benefit from using a UK dataset to understand how well their algorithm performs prior to UK launch and inform potential further development

Example Use Case: *Lancet Digital Health* **publication** *(2020) in which OMI-DB was used in the development of an AI algorithm for diagnosis of breast cancer in mammography*

Applying for Access

- OPTIMAM Mammography Image Database (OMI-DB) is open to applications from both academic groups and commercial organisations, please submit your research applications via the [online form](#)
  - The web form will alert the team of your interest, and they will send out a more detailed application form to be completed in full which will be reviewed by the [OPTIMAM Steering Committee](#)

- Access is subject to a data sharing agreement with Cancer Research Horizons and Royal Surrey NHS Foundation Trust

- Please find further details about the database at the Royal Surrey NHS Foundation Trust [OMI-DB website](#) or for licencing questions please reach out the CRH team