# Google Health mammography research FAQs January 2020

#### Do the results conclude anything about the differences between readers in the US and the UK?

No. There were no instances in which readers from the US and readers from the UK read the same set of cases. Each set has a unique demographic and pathologic mix, andthe screening programs differ significantly between the two countries, most notably in the screening interval (1-2 years in the US; 3 years in the UK). These differences make it difficult to compare reader performance between our US and UK datasets.

### Is the mammography dataset available to other researchers?

The US dataset was used under license for the current study, and is not publicly available at present. The UK dataset was compiled by Cancer Research UK, and researchers can apply for access <u>here</u>.

## Was there an external test set used for validation?

As described in the paper, we performed an experiment in which we trained the AI system using data from the UK alone, and then evaluated it on an unseen US test set. In this context, the model had no exposure to data from the US. Still, performance exceeded that of radiologists operating in clinical practice. We are excited by this evidence of the model's ability to generalize to different healthcare settings and will continue to pursue work to further demonstrate its effectiveness.

## How does this paper compare to prior work?

We appreciate the excellent work that has been done by many others (see <u>paper</u> for references), these papers have each contributed to the body of evidence for using AI to assist in breast cancer screening. However, we believe our study advances the field on several dimensions.

- Instead of drawing data from one site or location, we evaluated our algorithm on data from a large US academic medical center as well as multiple sites in the UK. Furthermore, we showed evidence of generalization of the models from one country to the other.

- We showed that our model's outputs compare favorably to thousands of interpretations made in routine clinical practice in the US and the UK, rather than a purely laboratory environment.

- We used a follow up period of 2 or 3 years after first screening mammogram, aligning with the screening intervals in the US and UK. This longer period has the advantage of including cancers that may have been missed by human eyes during the original screen.

- Finally, we quantified the AI system's localization ability, a critical feature for the interpretability of AI systems in a medical context.

## Does the AI system only find cancers that humans found?

Ground truth cancer status was determined on the basis of longitudinal followup for 2 and 3 years after screening. These intervals were chosen to encompass a subsequent round of screening in each country (the US and UK, respectively). As a consequence, the dataset includes cancers that may have been initially missed by the human readers. Such cancers could have been discovered at the next screening mammogram or presented symptomatically in the interim.

For the UK dataset, patient outcomes were determined through the <u>dataset's</u> linkage with the National Breast Screening System, ensuring that future cancer diagnoses would still be included even if some patients visited UK screening sites outside of those included in our study. It is through this means that we minimize the pitfalls of the "gatekeeper effect" that can crop up in retrospective studies.

#### How were the dicoms preprocessed?

The pixel data was extracted from the dicom files using the open source library <u>DCMTK</u>, respecting the window/level and lookup table values present in the header. Before feeding the images to the models, some rudimentary cropping and padding was performed, as described in our <u>supplementary methods</u>.

## Google Health mammography research FAQs January 2020

### How were the train/tune/test set sizes chosen?

The UK test set is a random sample of 10% of all women screened at two sites, St. George's Hospital and Jarvis Breast Screening Centre, between the years 2012 and 2015. Women from the US cohort were split randomly between train (55%), validation (15%) and test (30%). This scheme follows machine learning convention, but errs on the side of a larger test set to power statistical comparisons and include a more representative population.

The size of the reader study was selected based on time and budgetary constraints. The case list was composed of 250 negative exams, 125 biopsy-confirmed negative exams and 125 biopsy-confirmed positive exams. We sought to include sufficient positives to power statistical comparisons on the metric of sensitivity, while avoiding undue enrichment of the case mixture. Biopsy-confirmed negatives were included to make the malignancy discrimination task more difficult.

#### Why were some of the experience levels for the UK readers unknown?

All clinical readers represented in the UK dataset were part of the NHS breast screening programme, and so necessarily met minimum professional qualifications and experience. However, precise job title and annual case volume was available for just 18 of the 51 readers. This information is presented in Extended Data Table 7.