The Tyrer-Cuzick model is known to give imperfect risk predictions, meaning that the incidence of cancer in specific risk cohorts is not fully consistent with the risks assigned to that cohort. A 2018 study by Brentall *et al.* followed up individuals assigned 10-year risks using the Tyrer-Cuzick questionnaire and Volpara breast density measurement after ten years and measured the number of diagnosed breast cancer cases in each risk group and found discrepancies between the expected and observed numbers of cases[CITE BRENTNALL 2018 HERE]. In the updated analysis, we use the results of this study to correct for prediction error in the Tyrer-Cuzick model by calibrating parametric risk distributions to the expected and observed case numbers and then mapping between these two distributions.

The Tyrer-Cuzick model assigns each individual to one of 5 risk groups defined by bounds on that individual’s 10-year risk: <2%, 2-3%, 3-5%, 5-8%, and ≥8%. The supplementary material to Brentnall et al. 2018 provides the numbers of individuals in their study assigned to each risk group, stratified by age, along with the expected number of cases in each age-risk subgroup at time of risk stratification and the observed number of cases in each age-risk subgroup at the 10 year follow-up date. We denote the number of individuals in a given age group who are assigned to risk group by , with giving the total number of individuals in the sample, and denote the expected and observed number of cases in risk subgroup by and respectively.

We approximate the distributions of underlying and true risks underlying this data using beta distributions with parameters chosen by simulating the process that generates , , and . Given a beta distribution with parameters , we can simulate the generation of expected case numbers as follows

1. For each individual in a population of individuals, draw a 10-year risk ;
2. Reorder the 10-year risks so that for ;
3. For each individual draw a 0/1-valued cancer status ;
4. Generate synthetic expected case numbers by summing the number of positive cancer statuses in each successive block of individuals, i.e.
5. Carry out steps 1 to 4 1,000 times to get an estimate of the expectation of the expected number of cases in each risk group:
6. Calculate the root mean square error associated with the parameters :

Using the optim function from the stats package of the R Programming Language to minimise this root mean square error as a function of and , we identify optimal underlying expected risk distribution parameters . The same method applied to the observed case numbers gives us optimal parameters for the distribution of true risk.

To correct for prediction error in our synthetic population, we apply a quantile matching approach using the two fitted parameter sets, and . This approach relies on the assumption that the ordering of different individual’s risks is the same under the true underlying risk distribution as under the estimated risk distribution obtained from the Tyrer-Cuzick model, so that the true risk for every individual assigned to risk group is higher than that of any individual belonging to risk group and lower than that of any individual assigned to risk group . Given an individual with estimated lifetime risk , we define the corrected lifetime risk to be

so that an individual whose predicted risk is at the th percentile of the distribution fit to the expected incidence data will have a corrected risk at the th percentile of the distribution fit to the observed incidence data. This corrected risk follows the beta distribution fit to observed incidence data.

As well as 10-year risk, the Tyrer-Cuzick model estimates each individual’s lifetime risk of developing breast cancer. Because Brentnall *et al.* 2018 only includes data from a 10-year follow-up as opposed to lifetime tracking of outcomes, we can not generate a correction to lifetime risk using the same methodology as for 10-year risk. Instead, we generate a corrected lifetime risk   given an estimated lifetime risk by assuming that the same functional relationship holds between 10-year and lifetime risk based on both estimated and corrected values, i.e. and   for some function . To account for noise in the relationship between risk values, we approximate this function using Gaussian process regression, a Bayesian non-parametric machine learning method which infers a probability distribution of possible functional relationships between input data and output data. Once this distribution of possible functions has been estimated, it can be used to randomly generate new datapoints. We used the R Programming Language package GauPro to infer a Gaussian process from the estimated 10-year and lifetime risks in our synthetic population. For each individual in the synthetic population, we augmented their corrected 10-year risk with a corrected lifetime risk .

Based on our 10-year and lifetime risk estimate correction methods, we generated an augmented synthetic population identical in characteristics to that used in previous iterations of the model, with two additional fields specifying a corrected 10-year risk and corrected lifetime risk. The synthetic risk data includes individuals under 50 years old and between 50-60 years old. Separate correction models were fitted for these two age groups based on the age-stratified expected and observed case numbers reported by Brentnall *et al.*, and the appropriate correction model was chosen based on each individual’s age.