LUND
UNIVERSITY

A mixed clinicopathological and molecular proxy of homologous recombination deficiency in triple negative breast cancer

Elise Walford

**Master’s Degree Project in Bioinformatics, 30 credit
2021**

Department of Biology
Lund University

A mixed clinicopathological and molecular proxy of homologous recombination deficiency in triple negative breast cancer

Elise Walford (Lund University) BINP50

Abstract

Clinical models are increasingly employed in medical science as either diagnostic or prognostic aids. Machine-learning methods are able to draw links in large data that can be used to predict patient risk and allow more informed decisions regarding treatment and medication intervention. An advanced clinical predictor, HRDetect, can determine loss of homologous recombination-based repair pathways in patients with triple negative breast cancer, other breast cancers, and other cancers with high accuracy through the use of mutational signatures determined through whole genome sequencing. These patients respond well to treatment with targeted therapies, and the predictor is able to identify a far larger number of patients than would be identified using current clinical methods. The aim of this thesis was to predict the results of patients previously classified by HRDetect using only more clinically available data alone. The predictor was developed in SCAN-B data of triple negative breast cancer patients. The process utilised multiple imputation to handle missing values, modelling of continuous variables using restricted cubic splines, and sparse principal component analysis for dimensionality reduction. The model was internally validated using bootstrapping, adjusted to improve calibration and applied to an external breast cancer dataset for external validation. Interpretation of results was made difficult by large differences between the development and validation datasets, and the final model showed seemingly good discrimination but poor calibration. Further investigation of clinical relevance may be required.

Introduction

Advancements in sequencing technology have allowed genomes to be analysed with deeper coverage in faster times and at lower cost. This has lead to greater understanding of the genetic mechanisms beneath the surface in many fields, including cancer biology. Cancer is a leading cause of death, with up to one sixth of deaths attributed to it (1), with breast cancers alone accounting for up to 10% of cancer diagnoses (2). Deeper understanding of the genomic landscape that underpins these heterogeneous diseases has resulted in identification of new cancer driver and passenger mutations, and consequently, new therapeutic targets (3). Passenger mutations that have occurred prior to, during, and after cancer development can also be used to infer the "history" of cancer cells (4). Mutational signatures – specific patterns of mutations, for example, copy number changes and DNA rearrangements – can be closely associated with specific cancer types, tumour characteristics and with certain mutagenic processes or mutagen exposure (3,4,5). Taken altogether, these genomic patterns and relationships, understood from mutagenic signatures and genomic "scarring", can give vital information about a tumour, including prognostic indications, and can also inform what therapies may be beneficial to the patient.

One way in which breast cancers can be categorised is the expression patterns of three receptors: Estrogen, Progesterone and Human Epidermal growth factor type 2 (HER2) (6,7). Tumours expressing the Estrogen receptor (ER+) are receptive to certain hormonal treatments, while those that have HER2 amplification (HER2Amp) respond to trastuzumab (3,6,7). Triple-negative breast cancer (TNBC) lacks expression of all three receptors and accounts for up to 17% of breast cancers (6). It has a particularly poor prognosis but has been treated effectively with poly (ADP-ribose) polymerase (PARP) inhibitors (6). However, this treatment is only effective in tumours that have deficiency in homologous recombination (HR) -based repair pathways caused by a loss of function of both copies of BRCA1, BRCA2 or other related HR genes (8). Identification of this subgroup within TNBC is therefore important in informing clinical treatment for these patients. A recently developed predictive model, HRDetect, is able to do this; it can accurately identify patients with deficiency in these repair pathways using mutational signatures identified from whole genome sequencing (WGS), not only within TNBC but also within other breast cancer and other cancer types (9) This could allow identification of more patients who would benefit from treatment with PARP inhibitors.

Despite the advancements made in WGS, it is not standard practice in the clinic to use these technologies when making clinical decisions about patients. It is therefore interesting to investigate how well these groups can be identified from only those variables that are either currently routinely measured in the clinic or could become standard clinical practice in the near future. This was done by creating a predictive model using clinical and near-clinical data that had previously been assessed by HRDetect and determining how well this model could predict those patients identified by HRDetect as being deficient in HR-repair (HRD high).

Materials and Methods

Development Data:

Sweden Cancerome Analysis Network-Breast (SCAN-B) data (n = 237) was used in model construction. The data was collected between 2010 and 2015 from hospitals in Skåne. The data is of high quality, with few missing values in the main variables, and forms an incredibly representative population sample of TNBC in southern Sweden due to high levels of participation from patients (10). Clinical features measured included patient age, tumour size, status of tumour migration to lymph nodes (LN binary), cancer molecular subtype as assessed by different subtyping methods (PAM50 AIMS, PAM50 SSSP), biopsy cell content (assessed by a pathologist), tumour grade, and level of the cell proliferation marker Ki-67 amongst others. "Near-clinical" variables were those such as gene expression levels of BRCA1, hypermethylation status of the BRCA1 promoter, and known inactivation of BRCA1 and BRCA2.

External Validation Data:

Collected data from breast cancer tumours of all types across Europe, put together by the Breast Cancer Somatic Genetics Study (BASIS), which had also been previously analysed through WGS (4, 11) (n = 560) was used for assessing model performance in unseen data. While the data concerns WGS, clinical variables were also measured. The data was collected from many different institutions across Europe (including Iceland) and pooled. As a result, the external data is not necessarily a representative sample of the same population as the development data. Furthermore, different centres had different practices regarding variable measurement, resulting in some variables being missing from certain study locations. The methods used for measuring certain parameters also differed between the development data and external data, and there may be some biases introduced as a result of this. All data in both datasets had been previously assessed by HRDetect. Mutational data from the WGS for this data was also available.

Model Construction – Constraints and Considerations:

When determining the course that model building should take, one of the foremost considerations was the sample size of the development data. While the data is of excellent quality, the low number of datapoints limits the complexity of the model that can be created – if too many variables are included the model is likely to wind up overfit to the development data, leading to poor performance when exposed to new data (12). The model building method chosen was logistic regression, which is a standard modelling technique (14) and a suitable choice for modelling a binary outcome (HRDetect high / low) and one that supports a mix of categorical and continuous variables. Beyond this, studies have shown that it is an effective modelling solution for small datasets, in contrast to other model types and deep learning methods (15). The choice to use the BASIS data as an external validation set allowed the SCAN-B data to be fully used in model development, thereby avoiding inefficient train/test splits (12). Consideration was given to pooling the two datasets to allow a more robust model to be developed by giving access to a larger development sample size; however, ultimately, this avenue was not pursued as this would preclude external validation of the developed model.

Various studies have been carried out to determine the number of observations (“events”) that are required per variable included in the model (16, 17), (EPV). A ratio of 10:1 is commonly accepted as sufficient to avoid an overfit model (16-17), however other studies have suggested that ratios of 15:1

(18) or even 20:1 (15, 19) may lead to more robust and better fit models. This events per variable ratio is calculated based on the lowest frequency outcome in the data, in this case this is n = 98 (outcome = HRDetect low), suggesting that a model built in this data could support at most 9 variable "slots" using the lowest accepted ratio. However, different variables occupy different numbers of slots depending on their degrees of freedom. Factors take up n-1 slots, where n is the different factor levels the variable can take. Binary variables occupy a single slot (12). The number of slots for continuous variables varies depending on how they are modelled. With this limitation, some form of data reduction was necessary to maximise inclusion of explanatory parameters without overfitting the final model.

Models can also either be pre-specified (subject knowledge is used to select relevant variables to be included in the model) or data-based (variables are selected during the model construction based on their relationship with the outcome), and both options have their share of trade-offs that must be taken into consideration (20). Some data-based modelling techniques can have poor performance in small sample sizes (17), leading to inaccurate models. It is possible this may be mitigated through a sophisticated penalization approach such as lasso or the elastic net. Model pre-specification was ultimately chosen for this project, both to avoid potential pitfalls with data-based methods, and because predictors with association to the outcome are already known – i.e. there was subject knowledge to facilitate modelling decisions. When using pre-specification, variables should not be discarded from the model due to low significance (12).

Variable choice and handling:

Variables were chosen through pre-specification, as outlined. This involved selecting those that had known associations with HRDetect status and TNBC and treatment, for example age. While HRDetect is able to identify patients of all age groups who have the HRD phenotype, an early onset of cancer is known to be associated with this phenotype. Variables that had a very high degree of missingness or those that were deemed subjective were discarded, for example, cell content from tumour biopsy may be affected by which part of the tumour was captured. Other modelling studies have discarded variables with over 20% missing values (21). The final variables considered for model building were age, tumour size, tumour grade, LN binary, Ki-67, PAM50AIMS, BRCA1 gene expression, BRCA1 biallelic loss, BRCA2 biallelic loss and BRCA1 promoter hypermethylation. BRCA1 gene expression and tumour size both had skewed distributions and were log transformed before multiple imputation was carried out to normalise the distributions.

These chosen variables were then reduced to conserve “model slots”:

- PAM50AIMS classification was converted to a binary “basal / non-basal” variable. Basal morphology is associated with TNBC (6) and (loosely) with HRD phenotype. Other factor levels were much less common in the data, so this pooling of rare category levels seemed sensible.
- BRCA1/BRCA2 loss and BRCA1 promoter hypermethylation were converted into a single binary variable. A loss of function of any major gene in the HR-repair pathway should result in the same phenotype in terms of HR-repair, so all inactivation mechanisms can be considered as a single “suspected loss of pathway” variable.
- Continuous variables (age, tumour size, Ki-67 and BRCA1 gene expression) were modelled using restricted cubic splines with 3 knots. Ideally those variables thought to be most important would be given more knots, but due to the sample limitation all were given the minimum number of knots.

The use of restricted cubic splines avoids loss of information from categorising continuous variables (13) and can serve to relax the linearity assumptions required for variables (12). They also solve the problem of edge values in continuous data (12, 14). The number of model slots used by continuous variables is equal to the number of knots they are modelled with.

With this treatment, the variables in the model occupied a total of 16 slots (4 binary variables, and 4 continuous variables with 3 knots each), so further data reduction was still required to avoid overfitting. The restrictions also meant that interaction terms could not be included without the risk of creating a heavily overfit model. This was not thought to be a significant issue as there is evidence that inclusion of these terms often does not significantly improve a model (22).

Handling Missing Data:

Missing data distribution in both datasets was examined. There was little missing data in the development set within the variables chosen for inclusion into the model. Despite the low missingness, it was decided that missing values should be imputed, rather than discarding data with missing values. Use of complete cases (discarding of any data that contains missingness) is generally discouraged unless data is missing completely at random (MCAR) as it can result in a biased model in cases where missingness is not random but correlated with other significant values (26-29). The method chosen was multiple imputation, which was selected due to its high performance in simulation studies regarding the best handling of missing data (23). Multiple imputation was carried out using fully conditional specification using the MICE algorithm with predictive mean matching. This method is applicable to mixed categorical and continuous data types (26). Furthermore, it has been shown to be a forgiving method, particularly with respect to avoiding problems with misspecification of the imputation model, produces good results (24), and handles non-normal variable distributions well (30). The method involves creating a pool of donor values for the missing value based on other, similar data from the set, this results in only "real" values being chosen for replacement. Multiple imputation then generates separate datasets that use different replacement values drawn from the pool (24). Ten imputations were used for multiple imputation.

Data Reduction

Despite the variable handling steps undertaken, further data reduction was still required to adhere to the ratio rules for EPV. This was achieved through sparse principal component analysis (sPCA). This method resembles a combination of variable clustering and principal component analysis (PCA). Both these techniques can reduce high dimensional data (31); however unlike PCA, sPCA attempts to reduce component loadings to zero. This can make it an attractive alternative as it achieves data reduction while leaving the generated components more interpretable than those produced in PCA (32).

Like transformation, component calculation was blinded to the outcome, so components are based on the variance in the data they explain, and not on the variance in the outcome. The first six components explained a high proportion of variance in the data and were used as the final variables in the model. As the significance to the outcome was not measured, this was not thought to stray into data-based methods. By using sPCA, it was possible to reduce the dimensionality to six parameters, satisfying the EPV ratio rules outlined previously. This method should also solve collinearity in the variables (satisfying the assumption of variable additivity for logistic regression) as the method resembles a clustering step which is followed by a principal component analysis applied to the clusters (12). The drawback of this method is that, while the components are more interpretable than those of standard PCA (32), the overall interpretability of the model is reduced by the abstraction of the original variables.

Internal Validation and the naïve model:

It is important to validate a created model both internally and externally (17, 33-34). Internal validation can be thought of as validating the model building process itself. A created model will often be at least partially overfitted to its development data (34) and consequently perform less well in new data. Internal validation can correct the performance metrics of a model for development bias (35). In cases where external validation data is not available, internal validation methods are particularly important to provide an estimate of model performance without the need for creating unseen test data through data splitting (33). The bootstrap was chosen for internal validation, as it is a standard method that has shown good performance (36). The bootstrap repeats the model creation process in resampled datasets generated from the development data. It can also be used to calculate the "optimism" of model validation metrics and adjust them to reflect more accurate values (37-38).

It was also decided to use the bootstrap to calculate a global shrinkage factor that could be used to adjust the model coefficients to correct for/penalize overfitting to the development data (22). This method has been undertaken in other clinical models (21, 39). This is achieved by creating a first model, described here as the "naïve model". This naïve model followed the previously outlined steps using the full development data (n = 237):

1. Multiple imputation of missing values in the development data
2. Transformation of continuous variables using restricted cubic splines
3. Sparse principal component analysis
4. Logistic model created from the first six components

The naïve model was fitted from the mids object (generated using mice), so was generated from a pooling of all the imputed datasets.The process was then repeated within 10,000 bootstrap resamples (n = 237) that repeated all model building steps, culminating in the creation of a "boot model" for each resample. There was one introduced difference in the method: instead of fitting the boot models to all imputed datasets, they were fit individually in each imputation.

Studies that have combined multiple imputation with internal validation have generally shown that the preferable order of these operations is bootstrap → imputation (25, 40-41), so this step was carried out within the bootstrap resamples. By applying these boot models to the unseen "out-of-bag" data – the data that was not selected within each resample - the calibration of the model can be determined from a calibration curve and averaged over the many replicates. The averaged slope and intercept of these curves can then be used to adjust the naïve model's coefficients and intercepts (21-22, 39), correcting for the development process and model overfitting. The boot models were also applied to the original development data, to calculate the optimism of the C index in the performance metrics of the naïve model.

The combination of multiple imputation and the bootstrap presented some problems. Firstly, both the bootstrap and multiple imputation generate multiple datasets, so combining them multiplies the computational processing required. Furthermore, the use of the out-of-bag data required it to be imputed separately to the resampled data. Simulation studies that investigated combining imputation and bootstrapping with a validation in the out-of-bag suggested that imputation under these circumstances should be blinded to the outcome (25). Generally, multiple imputation should consider the outcome (27), so this was a pitfall of the method. However, with so little missing data, and many binary variables with heavily skewed class distribution this decision was unlikely to have much adverse effect. One concern in particular was that the out-of-bag data would be too small a validation set from which to draw meaningful conclusions. It was hoped that the many repetitions (10,000 bootstrap replicates x 5 imputations = 50,000 generated boot models) would balance out skewed results that this could cause.

The naïve model was adjusted based on the metrics obtained within the bootstrap, creating the final model. An overview of the full process can be seen in Figure 1. Simple comparative models were also briefly investigated in part to compare possible different approaches. All modelling steps were conducted with a preset seed.

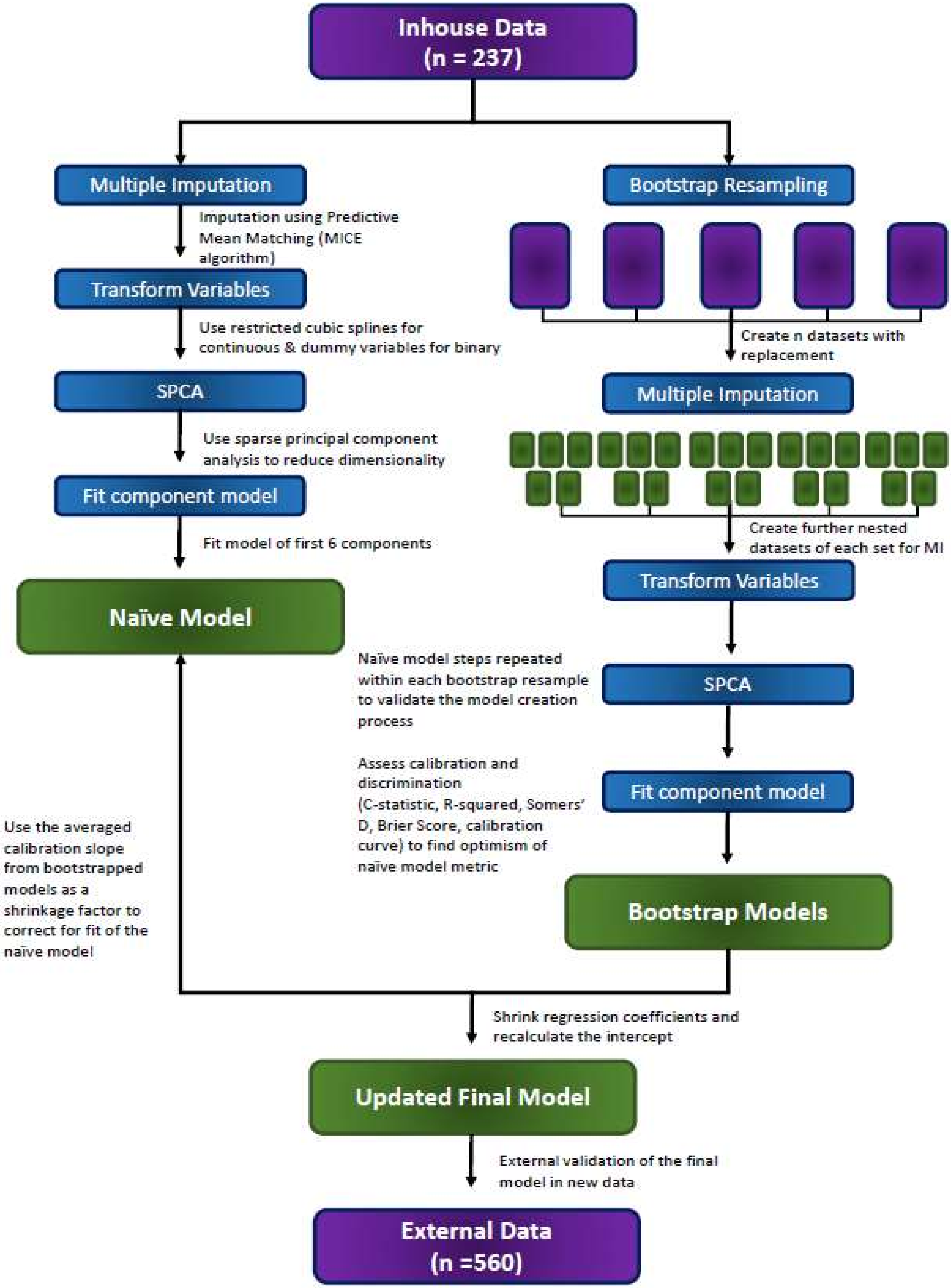

Inhouse Data
(n = 237)
Multiple Imputation
Imputation using Predictive Mean Matching (MICE algorithm)
Transform Variables
Use restricted cubic splines for continuous & dummy variables for binary
SPCA
Use sparse principal component analysis to reduce dimensionality
Fit component model
Fit model of first 6 components
Naïve Model
Bootstrap Resampling
Create n datasets with replacement
Multiple Imputation
Create further nested datasets of each set for MI
Transform Variables
SPCA
Fit component model
Bootstrap Models
Naïve model steps repeated within each bootstrap resample to validate the model creation process
Assess calibration and discrimination (C-statistic, R-squared, Somers' D, Brier Score, calibration curve) to find optimism of naïve model metric
Use the averaged calibration slope from bootstrapped models as a shrinkage factor to correct for fit of the naïve model
Shrink regression coefficients and recalculate the intercept
Updated Final Model
External validation of the final model in new data
External Data
(n =560)

Figure 1: Workflow overview of model development and internal validation

External Data & Validation of Final Model:

As previously mentioned, the external data was a somewhat flawed validation set for the clinical model, as measurement of clinical variables varied between institutions. Some variables, such as Ki-67 were not measured at all, while others were measured or not measured depending on institution. Tumour size, when measured, was categorised as a factor with 4 levels that translated to ranges. For the first two ranges (0-20mm & 20-50mm), the middle value of the range was assigned. The last two categories (which related to tumours beyond 50mm and those that had breached the basement membrane) were pooled due to low prevalence of the latter case. Tumours in these classes were treated as 55mm. While age was measured for all patients, discrete age was not given for participants above 80 years old. All patients who were over 80 were assigned an age of 85.

Multiple imputation was used to deal with the high degree of missing data in the external data, including a full imputation of the unmeasured Ki-67. This decision was to avoid using complete cases, which are not recommended if data is not missing complete at random (26-29). Multiple imputation has been shown to be effective even when imputing large amounts of data (42). The imputation was carried out using the development data as a framework, as this method was used in another study where a validation set had missing values (23). Imputation was performed by appending each external patient one by one to a previously imputed development set, imputing the new patient, then removing it and appending the next. The model and all requisite transformations were applied to the individual rows and a predicted probability between 0 and 1 was obtained for each imputed dataset. This corresponded to the predicted probability of the patient being HRDetect high. The final probability value taken was the average of probabilities from all imputed sets. Tumour size and BRCA1 gene expression were logged for both datasets prior to imputation.

The solution chosen had drawbacks. Since the datasets are not equivalent, it is possible that imputing in the development data may introduce bias. However, this option was preferable to imputing entirely within the external data, particularly when one variable had no measurements. Imputation becomes more difficult and less reliable the more missing values there are in the same set of observations, i.e. patients who have many missing variables are likely to have less accurate imputations than those missing few variables (29).

A validation using only a single imputation through predictive mean matching was also run alongside the multiple imputation, this resembled what one multiply imputed dataset might look like. This single imputation was used to examine imputed variable distributions and to compare some parts of analysis.

The external data was stratified by different groups and the model's predictive ability validated within these groups for both multiple and single imputation.

R Packages & Libraries:

- All analysis steps were performed in R 4.0.0 (43)
- tidyverse v. 1.3.0 including dplyr and ggplot2 used for tibbles and organization (44)
- readxl v. 1.3.1 used for importing datasets (45)
- naniar v. 0.6.0 for missing data plots and visualization (46)
- rms v. 6.0.1 for model building and validation (47)
- Hmisc v. 4.4-1 companion package to RMS for modelling (48)
- mice v. 3.11.0 for multiple imputation (49)
- boot v. 1.3.25 for creating the bootstrap (50-51)
- pcaPP v. 1.9-73 for sparse principal component analysis (52)

Results

Naïve Model & Internal Validation:

The missing data proportions from the development data can be seen in Figure 2A. The figure includes all potential clinical and near clinical variables, not just those that were used in the modelling process. Those variables with the highest levels of missingness were in most cases subjective variables which were not included in the model. The exception to this is heavy missingness in some other examined HR-repair genes, however it was decided to consider only the BRCA1 and 2 genes within the model as these are responsible for the majority of HRD cases and are more commonly measured.

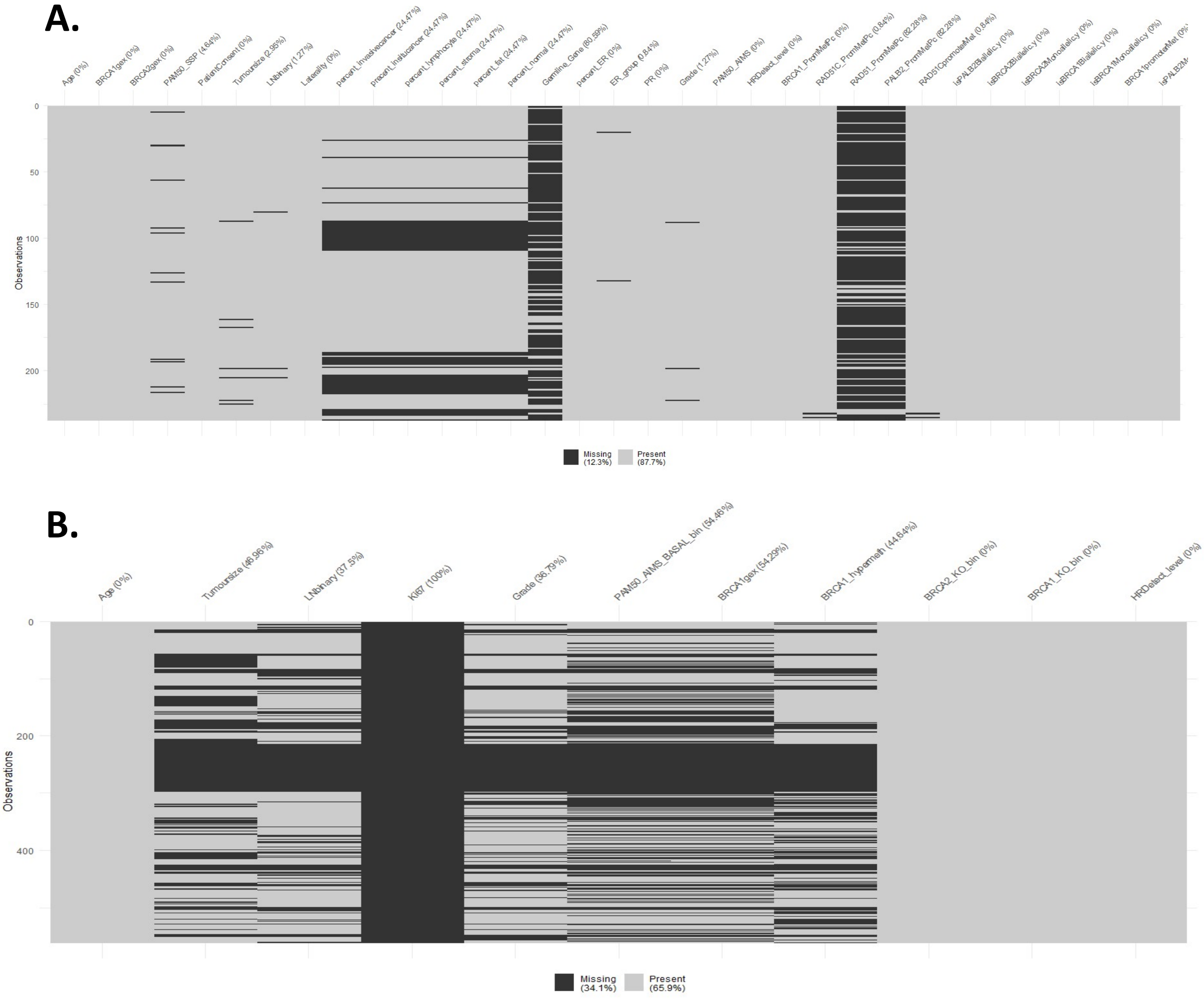


Figure 2: Missing data distributions – **A –** the development data, before variable reduction. **B** – the external validation data, showing only those variables that correspond to those used in model development.

The reported metrics from the naïve model development (Table 1) showed reasonably high scores for some discrimination metrics, such as C-index. However, good performance is expected in the development data. The metrics from the bootstrap of the model creation process are featured in Table 2. The optimism in the C statistic when boot models were applied to the full original development data was low, indicating the metrics should reflect the performance of the naïve model relatively accurately. In terms of the calibration slope (assessed in the out-of-bag data), the intercept is close to the desired value of 0, demonstrating good calibration-in-the-large. However, the slope is less than the ideal of 1, indicating some level of overfitting (53). Specifically, this indicates that high probabilities are overestimated and low probabilities are underestimated (54-55). The slope and intercept together were used to shrink the naïve model coefficients and recalculate the intercept to create the final model (21-22, 39, 56) (available in appendix). An alternative validation was trialled using the rms validate function (47), and the slope and intercept values from this matched well to those from the out-of-bag analysis. This may indicate that the repeated bootstrapping was able to correct for biases caused by the small validation sample on each attempt.

Table 1: Naïve model metrics from development

		Model Likelihood Ratio Test		Discrimination Indexes		Rank Discrim. Indexes	
Obs	237	LR chi2	157.01	R2	0.653	C	0.915
0	98	d.f.	6	g	3.331	Dxy	0.829
1	139	Pr (> chi2)	<0.0001	gr	27.976	gamma	0.829
max | deriv |	2e-08			gp	0.402	Tau-a	0.404
				Brier	0.114		

Table 2: Metrics of boot models averaged from 10,000 bootstrap resamples with 5 multiple imputation repeats

DEV_R2	DEV_CInd	Opt	OOB_CInd	OOB_Dxy	OOB_Brier	Cal_Int	Cal_Slope
0.6556	0.9141	-0.0133	0.8859	0.7719	0.1339	-0.0822	0.7845

A scree plot of the naïve model components showed that inclusion of the first six explained a cumulative 94% of the variance in the data. However, this was done while blinded to the outcome, so this does not correspond to an explained variance in the outcome. The components that were discarded were both single variable components of binary variables with very heavy class skew in the data, so this may be the reason for their low variance.

Comparative Models:

Two models that avoided sPCA were investigated. One used the same modelling procedure (multiple imputation followed by transformation within a bootstrap) but omitted the sPCA step for data reduction. The other used a previously imputed (single imputation) dataset using the RMS validate function to bootstrap the model for comparison. The degrees of freedom in both these models total 12, so they exceed the 10:1 ratio for events to the least frequent outcome. The model performance of both seemed broadly on par with the chosen strategy when considering validation metrics. Both did show a lower calibration slope, however, indicating greater overfitting. The value was not as different as expected (0.72 vs 0.78), where 1 is ideal.

External Validation:

An overview of the missing data proportions in the external data is available in Figure 2B. This figure shows only the equivalent parameters in the external data to those used in model development. Imputation of the external data was carried out while unblinded to the outcome, as this is the intended way for it to be performed. Brief investigations with a blinded imputation showed a generally worse performance. The degree of missingness was far higher than in the development data, and the distributions of variables may not be equivalent between the two sets. Imputed data distributions from a single imputation using predictive mean matching can be seen in the appendix.

Model performance was assessed on the external data as a whole and in various stratified subgroups. Assessment was made within a single imputed set and in multiply imputed sets. Performance was expected to vary substantially due to the considerable differences between the development data and the validation data. An overview of metrics in some of the investigated subgroups from multiple imputation is available (Table 3). The model appeared to show high discrimination across many of the groups examined in terms of the C-statistics and its related metric, Somers' Dxy. Despite this, the calibration measures show poor fit of the model in most cases. When considering performance in imputation groups, it is important to note that there were no "complete cases" in the external data due to the requirement to fully impute Ki-67. However, those that lacked only that variable have been loosely termed as "complete" for the purposes of comparison with those that required further imputation.

Table 3: Metrics from model performance in external data plus select stratified groups

		Validation Metrics					
		Dxy	**C (ROC)**	**R2**	**Brier**	**Intercept**	**Slope**
Overall		0.91281	0.95640	0.47398	0.09910	-1.28890	2.38416
Imputation level	complete cases	0.90991	0.95495	0.09215	0.08423	-2.40991	1.32825
	few imputations	0.94657	0.97329	0.61466	0.08154	-1.05197	2.87005
	many imputations	0.87940	0.93970	0.45668	0.11699	-1.02846	3.27739
Cancer subtype	TNBC all	0.84409	0.92205	0.64411	0.11722	-0.42570	2.27269
	TNBC complete	0.77778	0.88889	0.60279	0.13360	-0.77257	1.92426
	TNBC imputed	0.84818	0.92409	0.64705	0.11591	-0.39023	2.30264
	HER2 cases	0.49275	0.74638	-1.99206	0.14598	-1.60791	3.03651
	ER+ all	0.95462	0.97731	-0.00181	0.07868	-2.82695	1.76576
	ER+ complete	0.87211	0.93605	-0.52000	0.07968	-3.08298	1.30743
	ER+ imputed	0.98534	0.99267	0.16768	0.07813	-2.24831	7.14371
Age	<50 years	0.91852	0.95926	0.56436	0.11594	-1.24409	3.75441
	>50 years	0.89732	0.94866	0.30116	0.08720	-1.42271	1.89606

When considering the overall results, performance within data which had a high degree of missingness was worse than in those requiring fewer imputations. The more variables missing in a patient the more inaccurate the imputation can be (29), so this result seems understandable. The results within cancer subgroups seem to indicate a better performance in imputed data. However, within TNBC, the "complete" cases numbered only 12 patients, and therefore the effect of any poor results is heavily exacerbated by the small sample size. Within this data, there were two HRDetect high patients of more advanced age who did not have a known inactivating mechanism within BRCA1/2 despite having full data available. It is understandable that these outliers would not be assigned high probability by the model and affect the discrimination it displayed.

The missingness within different cancer subtypes is also not uniform, which somewhat confounds interpretation between the groups. A graphical overview of the model performance in data stratified by amount of imputation required overall, along with performance in different cancer subtypes with further division based on imputation can be seen in Figure 3. When considering the different breast cancer types collected, TNBC cancers make up around 30% of the external data. The bulk of the data corresponds to ER+ tumours (57%), with the remainder being HER2 amplified. Within TNBC, which would be expected to have the closest resemblance to the development data, there were 151 cases with missing data beyond Ki-67. Of these, 60 had at least 5 missing variables. HER2Amp tumour data had no "complete" cases, and of the 73 cases in the data, 62 had an exceedingly high degree of missingness. Imputation of the data within a fully TNBC dataset probably contributed to the poor performance in this group, and it showed the worst overall model metrics. ER+ tumours had many more complete cases, and a much lower proportion of cases that required many imputations within the same patient. This may be responsible for ER+ showing an apparently better performance in imputed cases between different cancer types – the imputation is made more unreliable in other cancer subtypes which require substantially more variables to be added. Further investigation into the ER+ complete cases showed that one patient (PD6042) had been classified as HRDetect low despite having a listed biallelic loss of BRCA2 in mutational information due to a nonsense mutation and loss of the alternative gene copy. The model's assigning this a high probability of HRDetect high makes sense in this context, and the reason for it being low is unknown, but could be due to an error in measuring the second allele loss.

The boxplots in Figure 3 show a general separation between the HRDetect low and HRDetect high classes in terms of probability assigned. However, further investigation shows that this effect seems to be primarily caused by one of the model's variables: the suspected loss of the HR repair pathway through inactivation/loss of function of BRCA1/2. The component in the model that comprises this variable has the highest coefficient.

When using a multiply imputed external dataset, PD6042 was the only example of an HRDetect low patient being assigned a probability score of over 0.7. However, in a singly imputed dataset, there were six such cases including PD6042. Missingness in the remaining five cases was very high, with very little data available beyond age and mutational status (but not hypermethylation) of BRCA1/2 recorded. The ages of the five patients tended to be on the lower side. It is likely that imputation of these cases in the development data incorrectly assigned promoter hypermethylation as being present. Younger age is associated with the HRD phenotype, so imputation using "similar" donors may have selected for this. This was probably further exacerbated by inaccuracy of imputation due to high missingness. By multiply imputing this data, the probabilities in which hypermethylation was imputed as present were modulated by those sets in which the donor value was an absence, bringing the overall probability down to below 0.7. The metrics from a singly imputed dataset are available in the appendix.

Investigation of HRDetect high patients that scored below the arbitrary 0.70 cut-off (32 total in multiply imputed external data, 36 in singly imputed external data) showed that these always lacked a known inactivation of HR-repair. Some were missing hypermethylation analysis and, while this was sometimes imputed as being present, it did not occur with enough frequency to result in a high probability. There are several other genes beyond BRCA1/2 involved in HR-repair. It is highly likely these patients are deficient in HR-repair, with an unknown mechanism; e.g. through biallelic loss of another HR-repair gene. This could be confirmed for one patient. PD4875 in TNBC had biallelic loss of BRIP1, which was recorded as another HR-involved gene. While monoallelic loss of other HR genes was noted for others, a clear-cut mechanism could not be identified. The probabilities assigned to these patients varied, and the level of missingness was, again, high. Those of more advanced age were assigned particularly low probabilities.

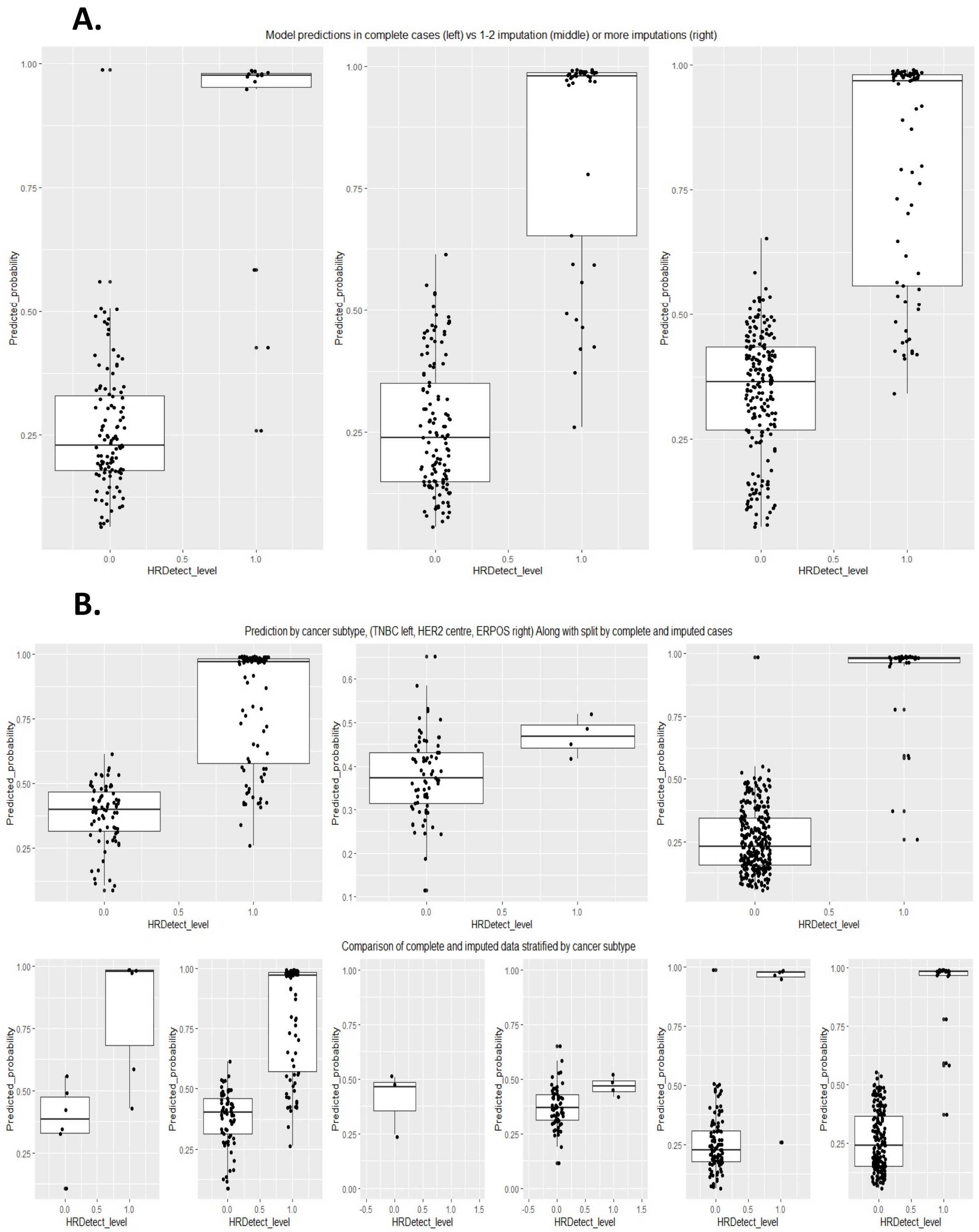


Figure 3: Box plots with scatter points of model performance in selected stratified groups in external data – **A – external** "complete" cases – only Ki-67 imputed (left), cases with few imputations (up to 3 including Ki67) (middle) and many imputations (>3 imputations) (right). **B** – stratification by breast cancer subtype overall performance (top row) and split based on "complete" cases and imputed cases (bottom row). TNBC (left), HER2Amp (centre), ER+ (right). Her2 amp "complete" cases are those with fewest imputations.

Overall, presence of a suspected inactivation of HR-repair resulted in a very high probability being assigned. Age and BRCA1 gene expression seemed to modulate the probability, with younger ages having a higher assigned probability and older ages often assigned very low probabilities, as would be expected from the known links between age and outcome. However, the major factor affecting probability was the inactivation of repair gene pathways. A waterfall plot (Figure 4) illustrates the assigned probabilities of HRDetect high patients in order, with colouration determined by whether an inactivation mechanism was known. This illustrates the strong effect of this variable on the model's assessment.

An investigation was made into how well the model could predict if this variable was entirely imputed (i.e. all values were removed in the external data and analysis was rerun). This was investigated using a single imputation with outcome, multiple imputation with outcome and multiple imputation where the external, but not the development, outcome was blinded. The results can be seen in Figure 5. In single imputation, loss of this pathway was correctly assigned to many patients, but was also assigned incorrectly to many HRDetect low patients, resulting in them receiving disproportionately high probabilities. Unblinded multiple imputation modulated this effect and showed much better distribution of the values. However, the blinded investigation shows that this is less clear in the absence of the outcome. The discrimination measured by the Cindex falls from 0.93 in unblinded multiple imputation to 0.77 in blinded imputation. The latter situation more closely represents the situation in the clinic – where the HRD phenotype may be unknown, and the outcome would always be unknown. The separation of the two groups effectively demonstrates the correlation between the other variables and the outcome and HRD phenotype, however. Without these associations multiple imputation would not perform in the way demonstrated.

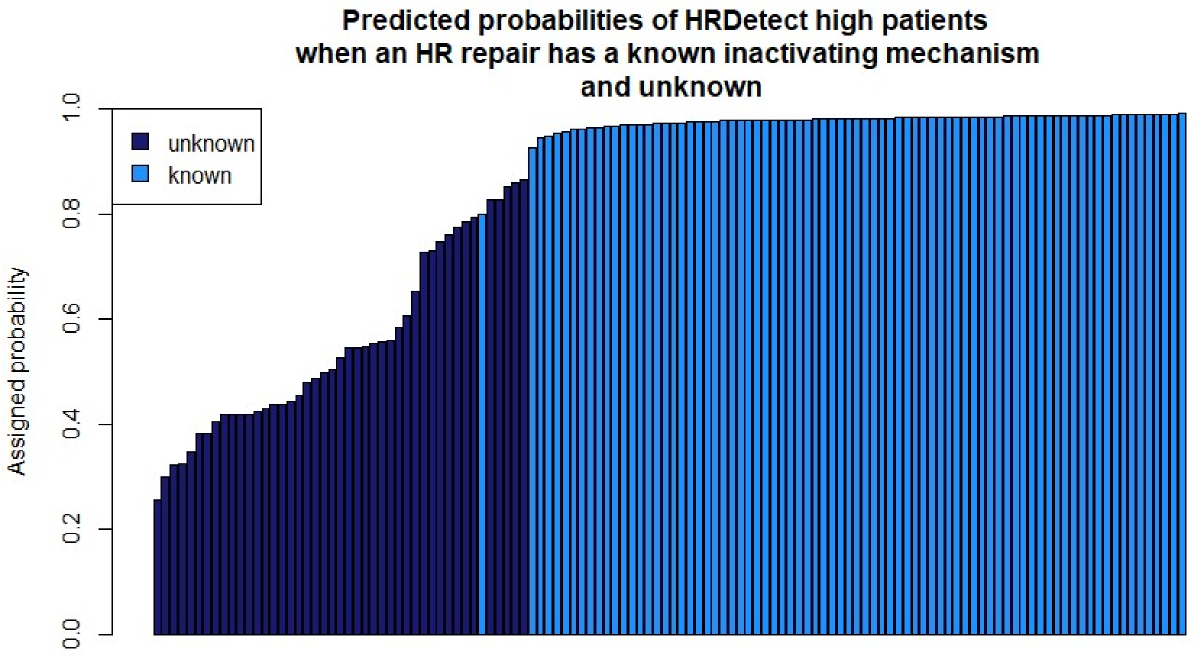


Figure 4: Waterfall plot showing model predicted probabilities of HRDetect high patients when status of BRCA1/2 are known vs unknown

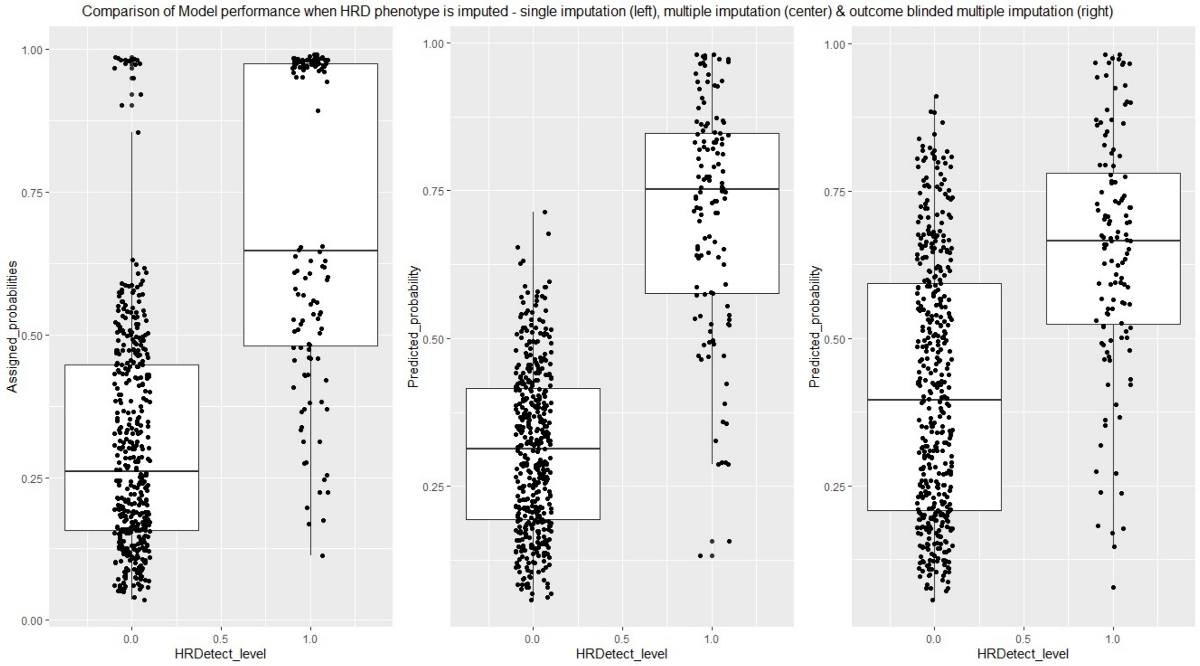


Figure 5: Box plot + scatter points of probabilities assigned by model in data with binary HR deficiency phenotype fully imputed through either single imputation, unblinded multiple imputation and blinded multiple imputation/ (For multiple imputation, model probability was averaged over 10 imputations)

Discussion

Overall, sample size was a major guiding force in modelling decisions. While the development data was of excellent quality, its size (n=237) introduced many constraints, resulting in concessions being made with the aim being to "spend" variable slots in the model in the most efficient way possible while adhering to ratio rules regarding events per variable. The largest impact of sample size was in the decision to pursue a component model, as proper modelling of the relevant continuous variables (using restricted cubic splines) constituted a large "expense" in terms of model variables. While this achieves heavy dimensionality reduction, it complicates usage of the model by requiring new data to undergo multiple transformation steps in order for the model to be applied. Furthermore, while sPCA could be seen as an improvement over variable clustering followed by regular PCA, interpretability of the model is still affected when compared to a model that considers the variables in their original states. There has been some debate regarding how suitable sPCA is for non-continuous variables, although regular PCA can be used for mixed variable types (31). If this application in non-continuous variables were an issue, the model could have selectively used sPCA for the continuous variables and maintained binary factors as dummy variables outside of sPCA. However, this would have resulted in eight parameters, thereby lowering the EPV ratio, although it would have remained within the confines of the 10:1 rule. Overall, it may have been preferable to find a solution that avoided the need for such extreme data reduction. The comparative models that omitted this step appeared to show slightly poorer calibration. A sophisticated penalisation method may have allowed inclusion of more variables while correcting for overfitting.

The internal validation of the model showed seemingly good discrimination, although there were some calibration issues. However, the use of the calibration mismatch as a global shrinkage factor is supposed to rectify overfitting. Nevertheless, this method is a somewhat crude measure as it applies evenly to all model coefficients. Penalisation of individual coefficients in the model allows for finer adjustment and may have been preferrable if different modelling methods were used. Penalisation can be implemented in some data-based model building methods such as LASSO and elastic net. Which method is best depends on the relative effects of the predictors in relation to each other (56). However, the size of the dataset in conjunction with these methods is still a concern. The implementation of multiple imputation within the bootstrap may not have been as good as it could have been, as it avoided proper pooling of the imputed datasets unlike the naïve model' development, and therefore may not have adhered correctly to Rubin's rules. However, the results were likely not significantly affected by this, as the variable included in the model had almost no missing values in the development data. Those that were missing were also almost all binary variables with very heavy class skew. The imputations for these would be expected to be the same – i.e. it would function no differently to single imputation. It was unclear whether the approach to external validation with multiple imputation, i.e. the averaging of model probabilities may also conflict with Rubin's rules regarding pooling from multiply imputed datasets. If this did turn out to be a problem, a deeper understanding of the methods used in the paper (23) which imputed using the development data may rectify it.

Another approach that could have been taken to mitigate the sample size constraints would have been to pool the BASIS and SCAN-B data – either for only the TNBC cases or the full data. However, this would have necessitated including cancer subtype as a variable, and the resultant model would then not be specific to TNBC. This would have expanded the sample size, possibly allowing for a more robust model to be created. Pooling of data in this way, followed by cross-validation that is stratified by the data collection institution has been proposed and used before, and is referred to as "internal-external cross-validation" (57-58, 39). This serves as a way to build a model from all available data, avoiding holding back data for validation, while still attempting to obtain a good estimation of future model performance. However, it is uncertain whether the vast differences in the way data was measured and collected between the two datasets may have made this unfeasible overall. It is unclear if the cross-validation being carried out by centre would rectify this, as validation would still be done partially between an all TNBC centre and centres of mixed cancer types. Imputation may also have posed a greater problem.

The external validation of the model was somewhat confounded by the differences between the development and test sets. A higher degree of missingness seemed to be associated with poorer overall model performance. This may have mostly been caused by the data in the HER2 subgroup, which was nearly entirely missing, and further compounded by the imputation using the development data as a basis. Since predictive mean matching selects donor values based on like data, the imputation may have been flawed as the two datasets were not equivalent; there were biases in the external data that were not present in the development data, and there were also other considerable differences between them beyond these. These differences occurred not only in the type of breast cancer assessed, but also in which variables were measured and how measurement was performed. The experimental procedures used to measure, e.g. hypermethylation in the development data were more sensitive than those used in the external data. Tumour size was also more biased in the external data towards larger tumours, where the development data was unbiased. These biases may have lead to inappropriate selection of the donor pool values in the imputation. Beyond this, predictive mean matching can become flawed when missingness exceeds a certain threshold (30). However, a comparison between a single imputation and multiple imputation using predictive mean matching seemed to show a modulating effect of multiple imputation, leading to better overall results than those obtained from a single imputation. A further concern is that missingness of variables may not have been random (MNAR), which has implications regarding the use of multiple imputation beyond those specific to predictive mean matching, which is designed for data that is missing at random (MAR).

Comparisons between cancer subtypes were made more difficult both by the difference in missingness between them and the sizes of the samples. When evaluating model performance within different degrees of imputation, many of the more complete cases belonged to ER+; a cancer subtype that the model had not seen before. Furthermore, performance in specific subtypes, such as HER2, may be affected by many missing cases in those types. Strangely, the model appears to perform better when few imputations are required than in complete cases, but this could be due to a small sample of complete cases which also contains outliers from the standard variable associations, such as in TNBC that disproportionately distort performance. Most of the more complete cases were in ER+ cancer which the model is unfamiliar with. This may also, in part, explain why the model appears to show unexpectedly good performance in ER+, as many imputations within the same patient can lead to poor values chosen, and the degree to which data was missing in ER+ was much lower than in TNBC and HER2. While the discrimination values were overall high, the calibration metrics show that the model is poorly fit. This may in part be caused by a much different proportion of HRDetect high cases in the validation data than in the development data. HRDetect high makes up approximately 60% of outcomes in the SCAN-B data, but only 22% in the BASIS data, of which 70% are in the TNBC. Overall, this makes the external set an imperfect validation set. A better validation set would be another iteration of SCAN-B, collected several years later. This would have been a test in a similar population, with variable collection carried out in the same way in terms of what is measured and how.

The comparison between single imputation and multiple imputation reinforced the benefits that multiple imputation provides according to prior studies. The averaging of probabilities across multiple imputations also helped to modulate high probabilities assigned when the imputation model assumed hypermethylation in a patient. This reduced the assignment of very high probabilities to HRDetect low patients with many missing variables that was seen in single imputation. This only worked because of the associations between variables and the outcome. The performance of the model may have been overstated by the use of the outcome to impute, because it correlated directly with the HRD phenotype variable. However, imputation is supposed to correct for what the data should be, and should not be blinded to the outcome, and the model was built using full data for this variable.

In terms of clinical usage of a model of this kind, the strong polarising effect of the HR-repair deficiency is problematic, as exemplified in the waterfall plot (Figure 4). While it is a "near-clinical" variable, it is not something routinely or necessarily easily measured in the clinic today, and a patient presenting with a known mutation in BRCA1/2 would likely be immediately assigned to PARP-inhibitor treatment without consultation of a model, simply on the basis of knowledge of the disease. While the model's discrimination metrics were reasonably good, in that generally HRDetect high are assigned higher probability than HRDetect low, it is unclear how much age and BRCA1 gene expression modulate the final probability without this variable. The imputation analysis which removed both outcome and knowledge of HRD phenotype gives the best representation of the clinic today (Figure 5, right). From this, it seems the model may have some ability to identify HRDetect high, if it involved a multiple imputation step in the development data, even without the HRD phenotype variable. The discrimination in this group was substantially lower, however and assigning a threshold that would balance capturing HRDetect high without also catching many HRDetect low may be challenging. Models should not necessarily be used to classify an individual. Instead, probability can be used to determine further action, for example, a deeper clinical investigation of those presenting with high predicted probability. However, further analysis would be needed to determine applicability to a clinical setting, for example with decision curve analysis. Model approximation could also be used to investigate model performance in the absence or presence of certain predictors; this could be used to better ascertain the predictive ability without the HR-repair deficiency variable.

References

1. American Cancer Society. Global Cancer Facts & Figures 4th Edition. Atlanta: American Cancer Society; 2018
2. Bray F, McCarron P, Parkin D. The changing global patterns of female breast cancer incidence and mortality. Breast Cancer Research. 2004;6(6).
3. Stratton, M., Campbell, P. & Futreal, P. The cancer genome. Nature 458, 719–724 (2009).
4. Nik-Zainal S, Morganella S. Mutational Signatures in Breast Cancer: The Problem at the DNA Level. Clinical Cancer Research. 2017;23(11):2617-2629.
5. Nik-Zainal, S., Davies, H., Staaf, J. et al. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. Nature 534, 47–54 (2016).
6. Foulkes WD, Smith IE, Reis-Filho JS. Triple-negative breast cancer. N Engl J Med. 2010 Nov 11;363(20):1938-48. doi: 10.1056/NEJMra1001389. PMID: 21067385.
7. Weigelt B, Geyer F, Reis-Filho J. Histological types of breast cancer: How special are they?. Molecular Oncology. 2010;4(3):192-208.
8. Glodzik D, Bosch A, Hartman J, Aine M, Vallon-Christersson J, Reuterswärd C et al. Comprehensive molecular comparison of BRCA1 hypermethylated and BRCA1 mutated triple negative breast cancers. Nature Communications. 2020;11(1).
9. Davies H, Glodzik D, Morganella S, Yates L, Staaf J, Zou X et al. HRDetect is a predictor of BRCA1 and BRCA2 deficiency based on mutational signatures. Nature Medicine. 2017;23(4):517-525.
10. Staaf, J., Glodzik, D., Bosch, A. et al. Whole-genome sequencing of triple-negative breast cancers in a population-based clinical study. Nat Med 25, 1526–1533 (2019) doi:10.1038/s41591-019-0582-4.
11. Nik-Zainal, S., Davies, H., Staaf, J. et al. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. Nature 534, 47–54 (2016) doi:10.1038/nature17676
12. Harrell, F. (2015). Regression modeling strategies (2nd ed.). New York: Springer.

13. Sauerbrei, W., Abrahamowicz, M., Altman, D., Cessie, S. and Carpenter, J., 2014. STRengthening Analytical Thinking for Observational Studies: the STRATOS initiative. Statistics in Medicine, 33(30), pp.5413-5432.

14. Sauerbrei, W., Perperoglou, A., Schmid, M., Abrahamowicz, M., Becher, H., & Binder, H. et al. (2020). State of the art in selection of variables and functional forms in multivariable analysis—outstanding issues. Diagnostic And Prognostic Research, 4(1). doi: 10.1186/s41512-020-00074-3

15. van der Ploeg, T., Austin, P., & Steyerberg, E. (2014). Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints. BMC Medical Research Methodology, 14(1). doi: 10.1186/1471-2288-14-137

16. Peduzzi, P., Concato, J., Kemper, E., Holford, T., & Feinstein, A. (1996). A simulation study of the number of events per variable in logistic regression analysis. Jou29rnal Of Clinical Epidemiology, 49(12), 1373-1379. doi: 10.1016/s0895-4356(96)00236-3

17. Steyerberg, E., & Vergouwe, Y. (2014). Towards better clinical prediction models: seven steps for development and an ABCD for validation. European Heart Journal, 35(29), 1925-1931. doi: 10.1093/eurheartj/ehu207

18. Chen, Q., Nian, H., Zhu, Y., Talbot, H., Griffin, M., & Harrell, F. (2016). Too many covariates and too few cases? - a comparative study. Statistics In Medicine, 35(25), 4546-4558. doi: 10.1002/sim.7021

19. Riley, R., Snell, K., Ensor, J., Burke, D., Harrell Jr, F., Moons, K., & Collins, G. (2018). Minimum sample size for developing a multivariable prediction model: PART II - binary and time-to-event outcomes. Statistics In Medicine, 38(7), 1276-1296. doi: 10.1002/sim.7992

20. Royston, P., & Sauerbrei, W. (2008). Multivariable model-building (1st ed., pp. 23-51). Chichester: Wiley.

21. Carmona-Bayonas, A., Jiménez-Fonseca, P., Lamarca, Á., Barriuso, J., Castaño, Á., & Benavent, M. et al. (2019). Prediction of Progression-Free Survival in Patients With Advanced, Well-Differentiated, Neuroendocrine Tumors Being Treated With a Somatostatin Analog: The

GETNE-TRASGU Study. Journal Of Clinical Oncology, 37(28), 2571-2580. doi: 10.1200/jco.19.00980

22. Moons K, Altman D, Reitsma J, Ioannidis J, Macaskill P, Steyerberg E et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): Explanation and Elaboration. Annals of Internal Medicine. 2015;162(1):W1.
23. Janssen, K., Vergouwe, Y., Donders, A., Harrell, F., Chen, Q., Grobbee, D., & Moons, K. (2009). Dealing with Missing Predictor Values When Applying Clinical Prediction Models. Clinical Chemistry, 55(5), 994-1001. doi: 10.1373/clinchem.2008.115345
24. Morris, T., White, I., & Royston, P. (2014). Tuning multiple imputation by predictive mean matching and local residual draws. BMC Medical Research Methodology, 14(1). doi: 10.1186/1471-2288-14-75
25. Wahl, S., Boulesteix, A., Zierer, A., Thorand, B., & van de Wiel, M. (2016). Assessment of predictive performance in incomplete data by combining internal validation and multiple imputation. BMC Medical Research Methodology, 16(1). doi: 10.1186/s12874-016-0239-7
26. Liu, Y., & De, A. (2015). Multiple Imputation by Fully Conditional Specification for Dealing with Missing Data in a Large Epidemiologic Study. International Journal Of Statistics In Medical Research, 4(3), 287-295. doi: 10.6000/1929-6029.2015.04.03.7
27. Sterne, J., White, I., Carlin, J., Spratt, M., Royston, P., & Kenward, M. et al. (2009). Multiple imputation for missing data in epidemiological
28. van Buuren, S., Boshuizen, H., & Knook, D. (1999). Multiple imputation of missing blood pressure covariates in survival analysis. Statistics In Medicine, 18(6), 681-694.
29. Horton, N. and Kleinman, K., 2007. Much Ado About Nothing: – a comparison of missing data methods and software to fit incomplete data regression models. The American Statistician, 61(1), pp.79-90.
30. Buuren, S. (2018). Flexible imputation of missing data (2nd ed., pp. 37, 52, 74, 75, 77, 82). Boca Raton: Taylor & Francis Group.
31. Lee S, Huang J, Hu J. Sparse logistic principal components analysis for binary data. The Annals of Applied Statistics. 2010;4(3):1579-1601.

32. Zou, H., Hastie, T., & Tibshirani, R. (2006). Sparse Principal Component Analysis. Journal Of Computational And Graphical Statistics, 15(2), 265-286. doi: 10.1198/106186006X113430

33. Steyerberg E, Harrell F. Prediction models need appropriate internal, internal–external, and external validation. Journal of Clinical Epidemiology. 2016;69:245-247.

34. Collins G, Reitsma J, Altman D, Moons K. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. BMJ. 2015;350(jan07 4):g7594-g7594.

35. Steyerberg, E., Eijkemans, M., Harrell, F. and Habbema, J., 2001. Prognostic Modeling with Logistic Regression Analysis: in search of a sensible strategy in small data sets. Medical Decision Making, 21(1), pp.45-56.

36. Steyerberg E, Bleeker S, Moll H, Grobbee D, Moons K. Internal and external validation of predictive models: A simulation study of bias and precision in small samples. Journal of Clinical Epidemiology. 2003;56(5):441-447.

37. Lankham I., & Slaughter M. (2020). Simple and Efficient Bootstrap Validation of Predictive Models Using SAS/STAT® Software. Proceedings of the SAS Global Forum 2020. Cary, NC: SAS Insititute Inc. Available: https://www.sas.com/content/dam/SAS/support/en/sas-global-forum-proceedings/2020/4647-2020.pdf

38. Miao Y., Stijacic Cenzer I., Kirby K. A., & John Boscardin W. (2013). Estimating Harrell's Optimism on Predictive Indices Using Bootstrap Samples. Proceedings of the SAS Global Forum 2013. Cary, NC: SAS Insititute Inc. Available: https://support.sas.com/resources/papers/proceedings13/504-2013.pdf

39. Hudda M, Fewtrell M, Haroun D, Lum S, Williams J, Wells J et al. Development and validation of a prediction model for fat mass in children and adolescents: meta-analysis using individual participant data. BMJ. 2019;:l4293.

40. Brand J, Buuren S, Cessie S, Hout W. Combining multiple imputation and bootstrap in the analysis of cost-effectiveness trial data. Statistics in Medicine. 2018;38(2):210-220.

41. Schomaker M, Heumann C. Bootstrap inference when using multiple imputation. Statistics in Medicine. 2018;37(14):2252-2266.

42. Madley-Dowd P, Hughes R, Tilling K, Heron J. The proportion of missing data should not be used to guide decisions on multiple imputation. Journal of Clinical Epidemiology. 2019;110:63-73.

43. R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/

44. Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686

45. Hadley Wickham and Jennifer Bryan (2019). readxl: Read Excel Files. R package version 1.3.1. https://CRAN.R-project.org/package=readxl

46. Nicholas Tierney, Di Cook, Miles McBain and Colin Fay (2020). naniar: Data Structures, Summaries, and Visualisations for Missing Data. R package version 0.6.0. https://CRAN.R-project.org/package=naniar

47. Frank E Harrell Jr (2020). rms: Regression Modeling Strategies. R package version 6.0-1. https://CRAN.R-project.org/package=rms

48. Frank E Harrell Jr, with contributions from Charles Dupont and many others. (2020). Hmisc: Harrell Miscellaneous. R package version 4.4-1. https://CRAN.R-project.org/package=Hmisc

49. Stef van Buuren, Karin Groothuis-Oudshoorn (2011). mice: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software, 45(3), 1-67. URL https://www.jstatsoft.org/v45/i03/.

50. Angelo Canty and Brian Ripley (2020). boot: Bootstrap R (S-Plus) Functions. R package version 1.3-25.

51. Davison, A. C. & Hinkley, D. V. (1997) Bootstrap Methods and Their Applications. Cambridge University Press, Cambridge. ISBN 0-521-57391-2

52. Peter Filzmoser, Heinrich Fritz and Klaudius Kalcher (2018). pcaPP: Robust PCA by Projection Pursuit. R package version 1.9-73. https://CRAN.R-project.org/package=pcaPP

53. Steyerberg E, Vickers A, Cook N, Gerds T, Gonen M, Obuchowski N et al. Assessing the Performance of Prediction Models. Epidemiology. 2010;21(1):128-138.

54. Huang Y, Li W, Macheret F, Gabriel R, Ohno-Machado L. A tutorial on calibration measurements and calibration models for clinical prediction models. Journal of the American Medical Informatics Association. 2020;27(4):621-633.

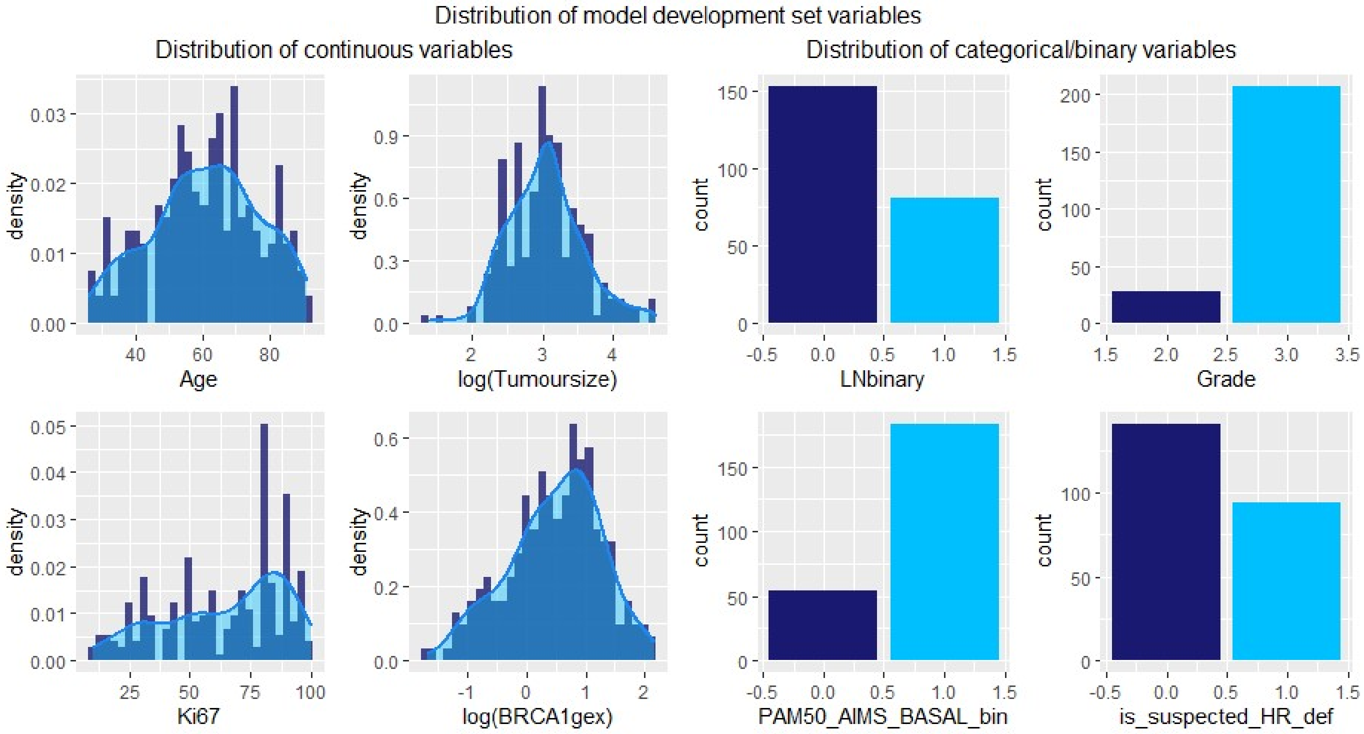
55. Stevens, R., & Poppe, K. (2020). Validation of clinical prediction models: what does the “calibration slope” really measure?. Journal Of Clinical Epidemiology, 118, 93-99. doi: 10.1016/j.jclinepi.2019.09.016

56. Steyerberg, E., 2019. Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating. 2nd ed. Springer.

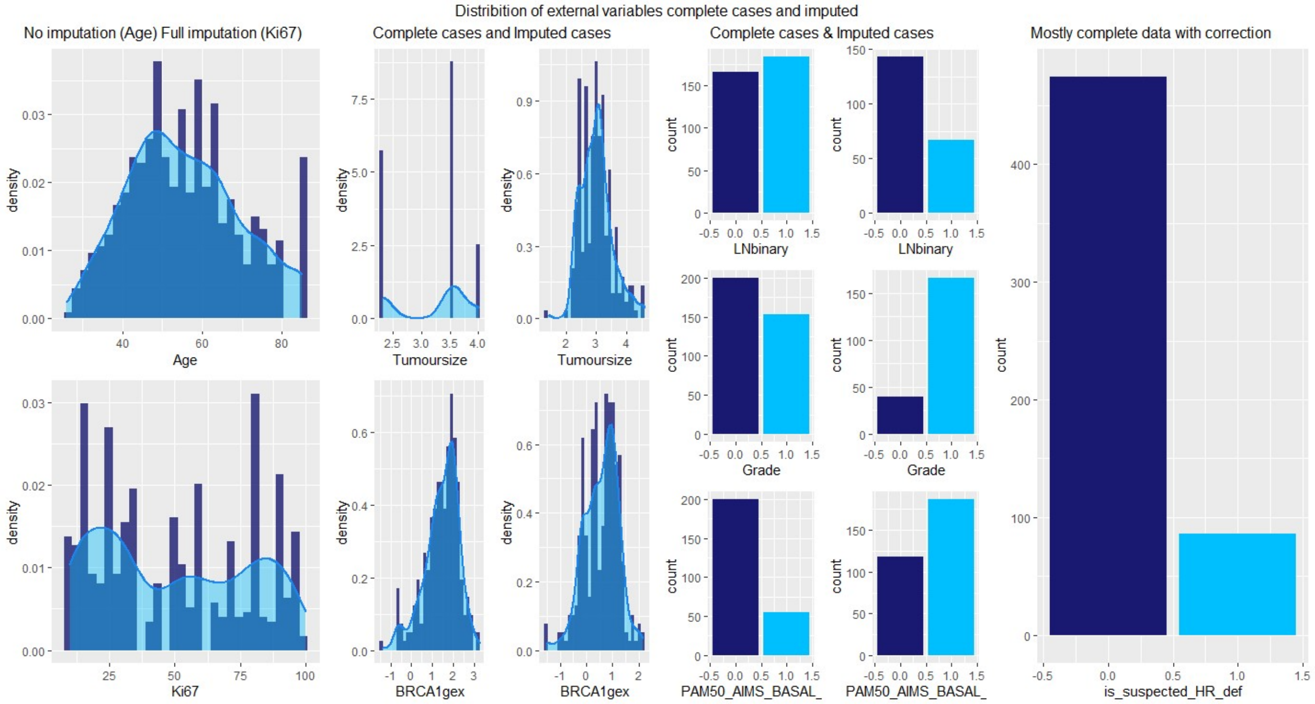
57. Steyerberg, E. and Harrell, F., 2016. Prediction models need appropriate internal, internal–external, and external validation. Journal of Clinical Epidemiology, 69, pp.245-247.

58. Royston P, Parmar M, Sylvester R. Construction and validation of a prognostic model across several studies, with an application in superficial bladder cancer. Statistics in Medicine. 2004;23(6):907-926.

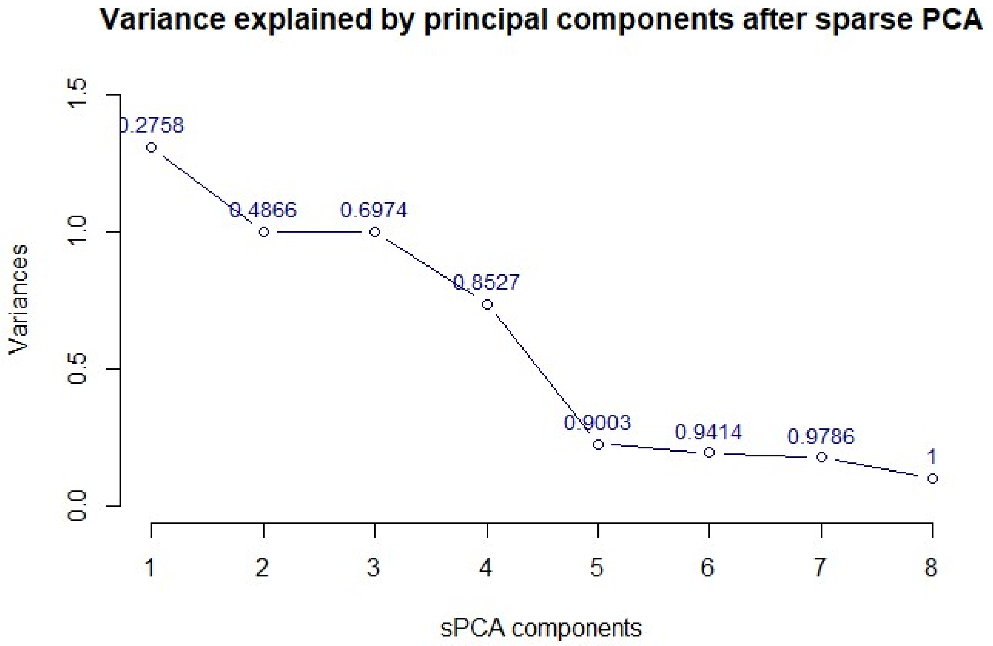
Appendix: Inhouse data distribution (continuous & categorical), BRCA1gex and Tumour size logged



Appendix: Distribution of variables in external data for (complete) age, fully imputed Ki-67 (left top and bottom), and paired distributions for known values and imputed values for tumour size, BRCA1 gene expression, and categorical variables. HRD was only partially imputed as some subfactors were complete. Imputation is single imputation using predictive mean matching



Appendix: scree plot of cumulative variance explained by components after sPCA (upper) and component loadings (lower)



Loadings:								
	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7	Comp.8
Age	0.682			0.730				
Tumoursize		1.000						
Lnbinary					1.000			
Ki67			1.000					
Grade								1.000
PAM50_AIMS_BASAL_bin							1.000	
BRCA1gex	0.728			-0.684				
is_suspected_HR_def						0.997		

Appendix: naïve model coefficient scores (left) and updated scores following shrinkage

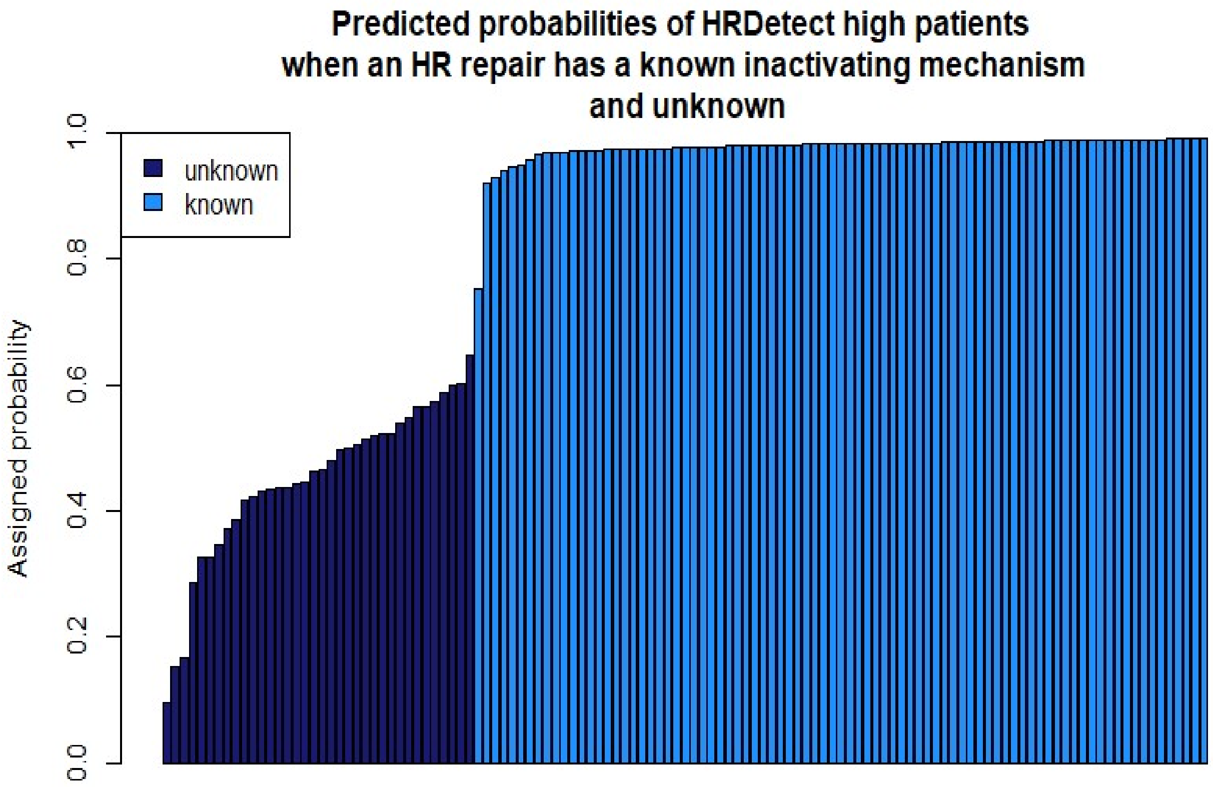
Naïve Model				
	Coef	S.E.	Wald Z	Pr(>|Z|)
Intercept	1.3771	0.4043	3.41	0.0007
Comp.1	-0.4165	0.2630	-1.58	0.1132
Comp.2	-0.1370	0.2129	-0.64	0.5200
Comp.3	-0.7145	0.2592	-2.76	0.0058
Comp.4	-0.5574	0.2979	-1.87	0.0613
Comp.5	-0.0919	0.4258	-0.22	0.8290
Comp.6	5.3945	1.1928	4.52	<0.0001

Adjusted (Shrunk) Model				
	Coef	S.E.	Wald Z	Pr(>|Z|)
Intercept	0.9980	0.4043	2.47	0.0136
Comp.1	-0.3267	0.2630	-1.24	0.2141
Comp.2	-0.1075	0.2129	-0.50	0.6138
Comp.3	-0.5605	0.2592	-2.16	0.0305
Comp.4	-0.4373	0.2979	-1.47	0.1421
Comp.5	-0.0721	0.4258	-0.17	0.8655
Comp.6	4.2319	1.1928	3.55	0.0004

Appendix: Validation metrics for model applied in single imputed external dataset (Imputation through predictive mean matching)

		Validation Metrics Considered					
		Dxy	C (ROC)	R2	Brier	Intercept	Slope
overall score		0.84827	0.92413	0.36998	0.11566	-1.73052	1.16281
imputation level	complete cases	0.83912	0.91956	0.13893	0.08091	-2.36607	1.20039
	few imputations	0.92799	0.96400	0.61574	0.08422	-1.38339	2.06919
	many imputations	0.75743	0.87871	0.22375	0.15146	-1.63430	0.97917
cancer subtype	TNBC all	0.74516	0.87258	0.57275	0.13852	-0.63387	1.13634
	TNBC complete	0.88889	0.94444	0.64767	0.11764	0.76006	9.86695
	TNBC imputed	0.73245	0.86623	0.56621	0.14018	-0.62376	1.10051
	HER2 cases	0.21905	0.60952	-4.96637	0.19843	-3.12104	0.06241
	ER+ all	0.88845	0.94423	-0.10152	0.08503	-2.90296	1.27868
	ER+ complete	0.72245	0.86122	-0.45330	0.07769	-2.99641	1.14277
	ER+ imputed	0.93175	0.96588	0.01875	0.08898	-2.86786	1.35687
Age	<50 years	0.82501	0.91250	0.42070	0.14247	-1.60211	1.08133
	>50 years	0.83266	0.91633	0.23404	0.09670	-1.82537	1.24399

Appendix: Waterfall plot showing model performance (probability) following only single imputation of validation data


Predicted probabilities of HRDetect high patients
when an HR repair has a known inactivating mechanism
and unknown
Assigned probability
unknown
known