Bayezian's Approach to Breast Cancer Survival Prediction: Leveraging AI and Machine Learning on Genomic Data While Navigating Pitfalls

Introduction:

Breast cancer is a heterogeneous and complex disease and cases are rising globally, with a reported 2.3 million new cases (Kashyap et al, 2022). Due to the complexity of the disease, it is difficult to treat with a blanket approach. Common therapies are dependent on the molecular subtype of breast cancer tissue, defined by specific genetic expression profiles. There are four common subtypes: luminal A and B, HER2+/ER-, and basal/ triple negative (Johnson et al, 2021). Luminal A is the most common subtype, accounting for ~70% of tumours. The basal subtype has the worst prognosis largely due to the lack of targeted therapies and its aggressive nature (Bhushan et al, 2021). For a full characterisation of these subtypes, refer to appendix A. Information on cancer subtype is being used to aid in the clinical decision pathway for treatments and prognosis.

Machine learning and artificial intelligence (AI) have extraordinary potential to support decision making in clinical practice by generating insights from large datasets. AI has already been utilised to aid imaging tools for breast cancer tumours for improved patient diagnosis, as well as aiding clinical practitioners in the decision process for treatment pathways (Bernardo et al, 2019; Corti et al, 2022). This potential is propagated by the adoption of electronic health records (EHR) from healthcare systems globally, with 96% reported use in the US in 2017, which contain information on diagnoses, treatments, and reported patient outcomes (Giordano et al, 2021). EHR have been utilised to create data warehouses, empowering researchers to train and test machine learning models on evolving real world data. These advancements allow for the adoption of personalised healthcare, so the needs of each individual are best accounted for, reducing mortality in patients.

Bayezian is combining data engineering techniques with regression machine learning models to predict breast cancer survival durations in patients. We used the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) database, containing information on genetic mutations, gene expression levels, tumour type, tumour stage, and treatment from 1,980 patients. In doing so, we are developing a tool that may help clinical practitioners decide on the best course of treatment to maximise prognosis in breast cancer patients, given personal medical history and genetic makeup. Bayezian sees this tool as a personalised approach to clinical decision making that puts people at the heart of patient care.

Despite advances in AI and its adoption into the clinical decision making process, Bayezian understands that many disadvantages remain. These stem from biases within the models, assumptions that some models make surrounding genomic data, and issues in data preparation. The risk associated with these, particularly in a clinical setting, is that the models do not meet the level of clinical viability that is demanded for patient care (Kelly et al, 2019). This paper will discuss these pitfalls in detail when applied to breast cancer survivability predictions. Additionally, it discusses how Bayezian is navigating these pitfalls and incorporating a robust approach to AI driven clinical decision making.

Distributional differences:

Within a dataset, disparities in distributions emerge when distinct variations in measured values coexist. Within genomic data, this phenomenon can arise from varying expression profiles of different biological structures, as demonstrated within the gene expression profiles of the breast cancer subtypes; refer to appendix A. In turn, challenges arise with maintaining consistent distributions between test, train, and validation datasets, as well as producing accurate predictions. Haque et al (2012) showed that HER2+/ER- and luminal B have a two-fold increase of mortality compared to luminal A, exemplifying how these distribution patterns have a direct impact on patient survivability outcomes. Therefore, it is imperative that these distributions are labelled with the associated explanatory variable, so the model can effectively account for these nuances. Our model retains the cancer subtype in order to account for this and increase the accuracy of our predictions.

Likewise, the presence of missing data can also contribute to distributional issues, due to the omission of data points lying between distributions. Additionally, data may be omitted within distributions, resulting in underrepresented groups. These scenarios diminish the statistical power of the model and amplifies its risk of bias (Marino et al, 2021). The intricacies of managing missing data are indeed challenging, and a universal approach to dealing with it does not exist. It is best practice to use complete datasets, although even the largest datasets are bound to have a degree of missingness. Our dataset boasts a completeness of over 95%, but a fraction of missingness persists. Removing outliers eliminated some of the missing data and an interpolation method was employed to fill in the remaining gaps. We used this over replacement methods that can result in skewed data. One method is to replace missing values with zeros, but this can cause the data to skew towards zero, altering its distribution and variance. Another relies on replacing data with random values, but unless the data is normally distributed, random groups of data may be artifactually inflated. Equally, imputation methods may inflate the mean group, causing a narrower distribution. Our method replaces missing values with an average of the value pre/proceeding it.The pre/ post- manipulated distributions were visualised for approximate equivalence to mitigate the marginal risk of bias. For a full visualisation of outliers and missing data, refer to appendices B and C, respectively. In turn, we've engineered a dataset that can be utilised to train a robust model for a reliable clinical calling.

Confounding data and overfitting:

Single genes can be involved in multiple processes and have many interacting partners (Stoney et al, 2018). Therefore, a gene may show correlation to a process, but is in fact regulated by another gene that is directly involved and the correlation observed is a confounding factor. Venet et al. (2011) showed there was a significant correlation between genes linked to social defeat obtained from mice brains and breast cancer survival outcome, demonstrating this concept. Whilst these may give accurate predictions, these may not be used to aid important decisions on patient care, such as treatments that may target specific genes involved in breast cancer. Equally, retaining these confounding variables can cause overfitting due to an overly complex dataset, so it is best practice to only retain variables directly involved in the process the model is trying to predict outcomes on.

In order to determine the most important features, Bayezian has drawn upon their domain expertise to distinguish the confounding variables within our dataset. Patients harbouring the BRCA1 mutation exhibit a significant reduction in their 10-year survival rate, as demonstrated by Huszno et al. (2019). Consequently, this mutation is associated with an elevated mortality rate in comparison to individuals without it. It has been determined that the synergistic use of anthracycline and taxane reduced mortality by 46% in mutation carriers (Godet and Gilkes, 2011). Here lies an example of how specific treatments may increase a patient's survival outcomes and vital information to include in a model such as ours, rather than confounding variables. Where it is difficult to characterise the confounding variables, particularly where data contains genes of unknown function, statistical methods may be employed to deal with this.

Addressing this within our dataset, we applied principal component analysis as a robust approach. This method serves the dual purpose of data preprocessing and feature reduction. It combines variables and weights them according to the features that explain the majority of the variance. Consequently, this strategy retains the bulk of the valuable information while simultaneously streamlining the dataset. We retained 537 principal components, which describes 80% of the variance and reduces the dataset by almost a third. This also generates uncorrelated variables to mitigate the confounding effects, as elucidated by Karamizadeh et al. (2013), when applied to making predictions on breast cancer prognosis.

Correlated data:

In machine learning regression models, the assumption of feature independence often falls short when dealing with genomic data, resulting in inaccuracies within model predictions due to correlated variables. Genes involved in the same process regulate each other, determining their respective expression levels. This means that these genes share interdependence. For example, increased BRCA1 expression, a gene involved in tumour suppression, is produced by elevated levels of the p63 gene when there is a break in the DNA (Crawford et al, 2010). A BRCA1 mutation will halt this process as the gene no longer functions as expected.This carries the risk of an individual being diagnosed with breast cancer, and carriers see a 72.5% diagnosis rate (Momozawa et al, 2022). With over 400 gene variants in the dataset, the likelihood of interdependence is high.

Bayezian has implemented XGBoost in our tool, harnessing the power of a boosted decision tree model adept at managing the complexities of multicollinearity. It uses a weak learner approach that relies on a simplified model, weighting the best performing features higher than underperforming ones. Co-linear variables notoriously result in models underperforming, so these would be weighted at or close to zero. Such information is iteratively fed into the next weak learner to increase its predictive power, until the error of the overall model is as close to zero as possible. The use of XGBoost on genomic data was corroborated by Prastyo et al (2020), who used machine learning algorithms to detect and classify breast cancer tumours. XGBoost performed the best with an F1-score of 98.51%. In our model, XGBoost performed best, seeing an improvement of 49.1 months in the root mean squared error (RMSE), compared to that of a standard linear regression model. For a full synopsis of models employed, refer to appendix D.

Conclusions and future outlook:

The incorporation of AI into the analysis of genomic data introduces challenges that complicate the construction of effective models. Furthermore, when deploying these models within clinical settings, as envisioned in our proposal, their accuracy and robustness become paramount. Bayezian recognizes the profound impact that decisions made regarding healthcare can have on individuals' lives. Consequently, the tools employed to facilitate such decisions must exhibit unwavering reliability. In this endeavour, we have demonstrated a comprehensive understanding of the potential pitfalls associated with AI in genomics, while also showcasing strategies to develop resilient tools for analysing breast cancer survival durations.

At present, our most advanced model achieves a RMSE of 55 months. While our application has yet to attain a clinically viable standard, it is evident that we have successfully navigated the complexities of data processing and model application, accounting for the intricacies inherent in genomic data. Notably, our dataset encompasses a substantial volume of information, posing a challenge for conventional machine learning models to effectively extract meaningful signals from the surrounding noise. Recognising the need to retain as much explanatory power as possible, we are inclined to explore the potential of deep learning algorithms. These techniques, characterised by multiple layers of data processing, hold the promise of enhanced accuracy in prediction by efficiently discerning the signal amidst the noise.

In essence, Bayezian is dedicated to the development of a pivotal tool that holds the potential to extend the lives of breast cancer patients. We embrace the multifaceted challenges posed by AI in genomics and remain committed to harnessing the power of advanced algorithms to surmount them, ultimately delivering a tool that empowers clinicians with dependable insights for improved patient outcomes.

References:

Bhushan, A., Gonsalves, A. and Menon, J.U. (2021). Current State of Breast Cancer Diagnosis, Treatment, and Theranostics. *Pharmaceutics*, [online] 13(5), p.723. doi:https://doi.org/10.3390/pharmaceutics13050723.

Bizzo, B.C., Almeida, R.R., Michalski, M.H. and Alkasab, T.K. (2019). Artificial Intelligence and Clinical Decision Support for Radiologists and Referring Providers. *Journal of the American College of Radiology*, 16(9), pp.1351–1356. doi:https://doi.org/10.1016/j.jacr.2019.06.010.

Corti, C., Cobanaj, M., Marian, F., Dee, E.C., Lloyd, M.R., Marcu, S., Dombrovschi, A., Biondetti, G.P., Batalini, F., Celi, L.A. and Curigliano, G. (2022). Artificial intelligence for prediction of treatment outcomes in breast cancer: Systematic review of design, reporting standards, and bias. *Cancer Treatment Reviews*, 108(1), p.102410. doi:https://doi.org/10.1016/j.ctrv.2022.102410.

Giordano, C., Brennan, M., Mohamed, B., Rashidi, P., Modave, F. and Tighe, P. (2021). Accessing Artificial Intelligence for Clinical Decision-Making. *Frontiers in Digital Health*, [online] 3, p.645232. doi:https://doi.org/10.3389/fdgth.2021.645232.

Godet, I. and Gilkes, D.M. (2017). BRCA1 and BRCA2 mutations and treatment strategies for breast cancer. *Integrative cancer science and therapeutics*, [online] 4(1). Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5505673/.

Haque, R., Ahmed, S.A., Inzhakova, G., Shi, J., Avila, C., Polikoff, J., Bernstein, L., Enger, S.M. and Press, M.F. (2012). Impact of Breast Cancer Subtypes and Treatment on Survival: An Analysis Spanning Two Decades. *Cancer Epidemiology Biomarkers & Prevention*, 21(10), pp.1848–1855. doi:https://doi.org/10.1158/1055-9965.epi-12-0474.

Huszno, J., Kołosza, Z. and Grzybowska, E. (2018). BRCA1 mutation in breast cancer patients: Analysis of prognostic factors and survival. *Oncology Letters*, 17(2). doi:https://doi.org/10.3892/ol.2018.9770.

Johnson, K.S., Conant, E.F. and Soo, M.S. (2020). Molecular Subtypes of Breast Cancer: A Review for Breast Radiologists. *Journal of Breast Imaging*, 3(1). doi:https://doi.org/10.1093/jbi/wbaa110.

Karamizadeh, S., Abdullah, S.M., Manaf, A.A., Zamani, M. and Hooman, A. (2013). An Overview of Principal Component Analysis. *Journal of Signal and Information Processing*, [online] 04(03), pp.173–175. doi:https://doi.org/10.4236/jsip.2013.43b031.

Kashyap, D., Pal, D., Sharma, R., Garg, V.K., Goel, N., Koundal, D., Zaguia, A., Koundal, S. and Belay, A. (2022). Global Increase in Breast Cancer Incidence: Risk Factors and Preventive Measures. *BioMed Research International*, [online] 2022, pp.1–16. doi:https://doi.org/10.1155/2022/9605439.

Kelly, C.J., Karthikesalingam, A., Suleyman, M., Corrado, G. and King, D. (2019). Key challenges for delivering clinical impact with artificial intelligence. *BMC Medicine*, [online] 17(1). doi:https://doi.org/10.1186/s12916-019-1426-2.

Marino, M., Lucas, J., Latour, E. and Heintzman, J.D. (2021). Missing data in primary care research: importance, implications and approaches. *Family Practice*, 38(2), pp.199–202. doi:https://doi.org/10.1093/fampra/cmaa134.

Momozawa, Y., Sasai, R., Usui, Y., Shiraishi, K., Iwasaki, Y., Taniyama, Y., Parsons, M.T., Mizukami, K., Sekine, Y., Hirata, M., Kamatani, Y., Endo, M., Inai, C., Takata, S., Ito, H., Kohno, T., Matsuda, K., Nakamura, S., Sugano, K. and Yoshida, T. (2022). Expansion of Cancer Risk Profile for *BRCA1* and *BRCA2* Pathogenic Variants. *JAMA Oncology*, 8(6), p.871. doi:https://doi.org/10.1001/jamaoncol.2022.0476.

Prastyo, P.H., Paramartha, I.G.Y., Pakpahan, M.S.M. and Ardiyanto, I. (2020). Predicting Breast Cancer: A Comparative Analysis of Machine Learning Algorithms. *Proceeding International Conference on Science and Engineering*, 3, pp.455–459. doi:https://doi.org/10.14421/icse.v3.545.

Stoney, R.A., Robertson, D.L., Goran Nenadic and Schwartz, J.-M. (2018). Mapping biological process relationships and disease perturbations within a pathway network. *npj systems biology and applications*, 4(1). doi:https://doi.org/10.1038/s41540-018-0055-2.

Venet, D., Dumont, J.E. and Detours, V. (2011). Most Random Gene Expression Signatures Are Significantly Associated with Breast Cancer Outcome. *PLoS Computational Biology*, 7(10), p.e1002240. doi:https://doi.org/10.1371/journal.pcbi.1002240.

Appendices:

Appendix A:

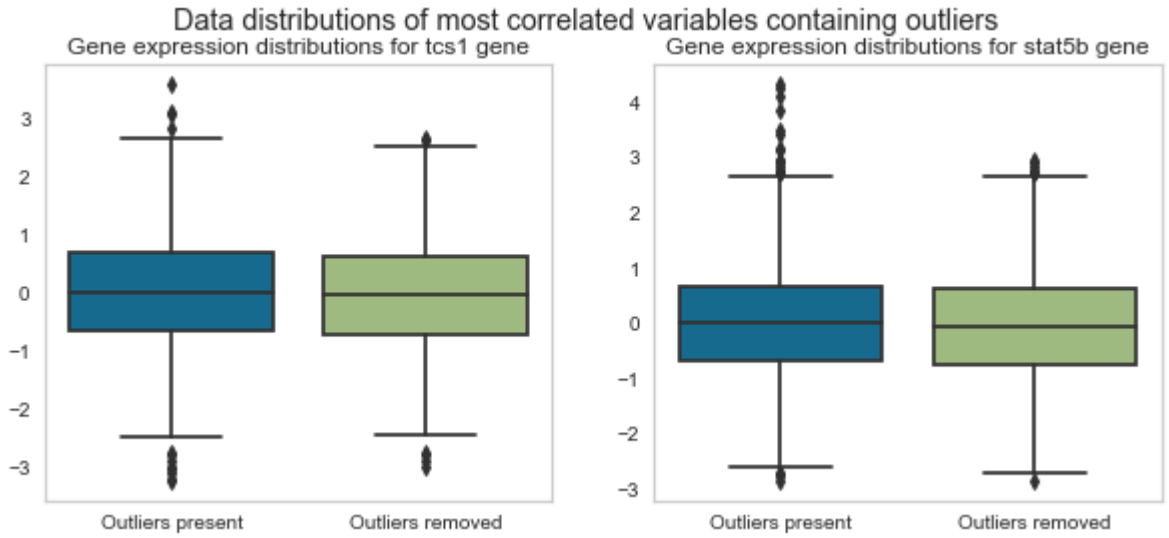
Table 1:

Subtype	Gene Expression Profile	Prognosis
Luminal A	• Estrogen receptor + (ER+) • Human epidermal growth factor - (HER2-) • Low levels of protein KI-67	Excellent prognosis due to low proliferation rate and existence of targeted therapies.
Luminal B	• Estrogen receptor + (ER+) • Human epidermal growth factor + or - (HER2+/-) • Elevated levels of protein KI-67	Good prognosis due to targeted therapies, but slightly worse than luminal B due to elevated protein KI-67 causing faster cell proliferation and aggressiveness.
HER2+/ER-	• Estrogen receptor - (ER-) • Human epidermal growth factor + (HER2+)	Good prognosis due to targeted therapies, although it is associated with higher proliferation rates than luminal cancers, making it more aggressive.
Basal/ Triple negative	• Estrogen receptor - (ER-) • Progesterone receptor - (PR-) • Human epidermal growth factor - (HER2-)	Worst prognosis due to no availability for targeted therapies and aggressive nature.

A full characterisation of breast cancer subtypes and their associated prognosis. Each subtype expresses different proteins that may be targeted for targeted hormonal therapies. This may not be applied to the basal subtype due to the lack of expression of these proteins (Bhushan et al, 2021; Johnson et al, 2021).

Appendix B:

Fig 1:

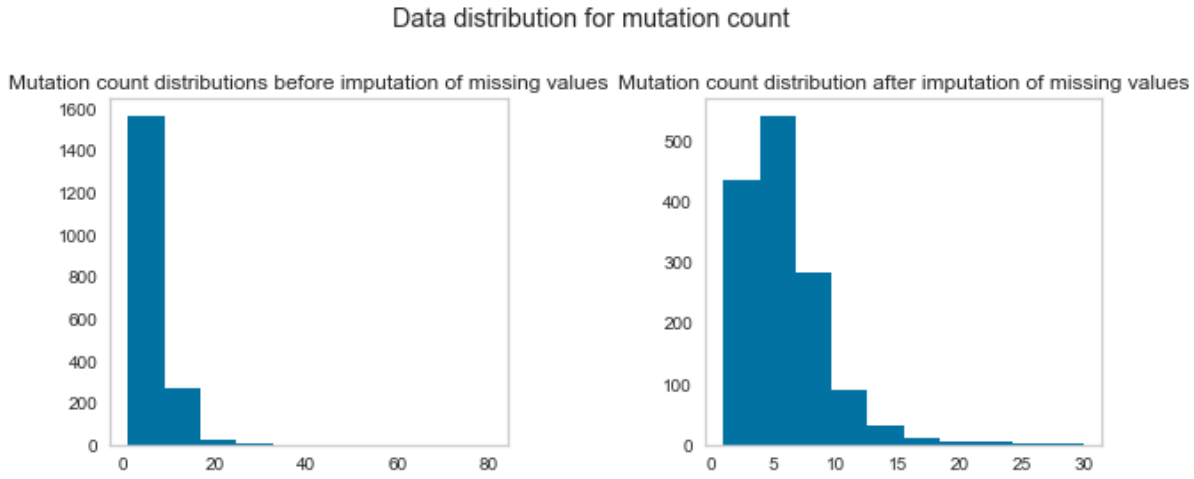


A schematic indicating outliers in gene expression values for *stat5b* and *tcs1* gene expression levels. The outliers are shown before and after removal. Outliers were removed where they had a z-score greater than three, retaining 95% of the data. Distributions retained relative equivalence, indicating success in outlier removal.

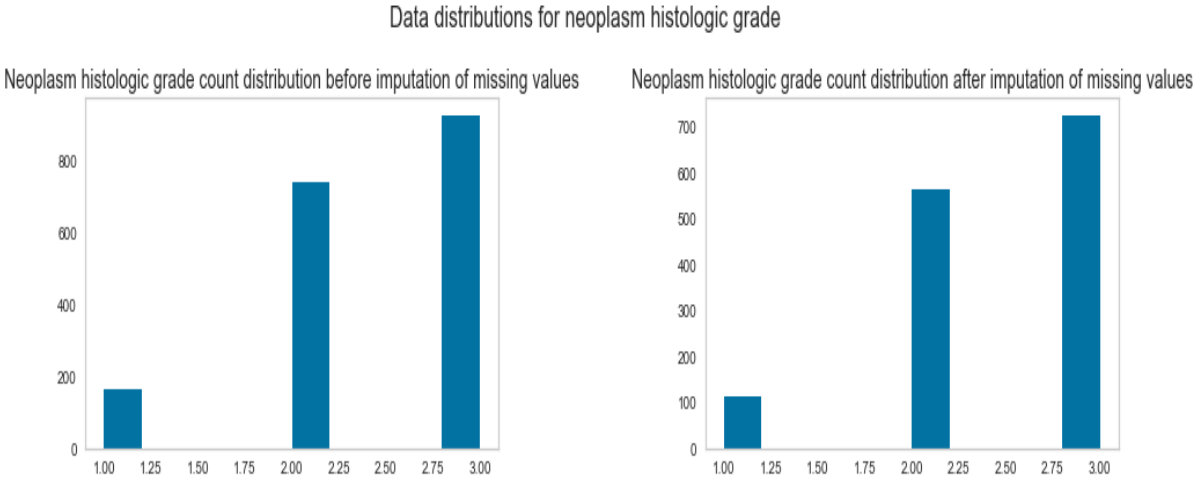
Appendix C:

Fig 2:

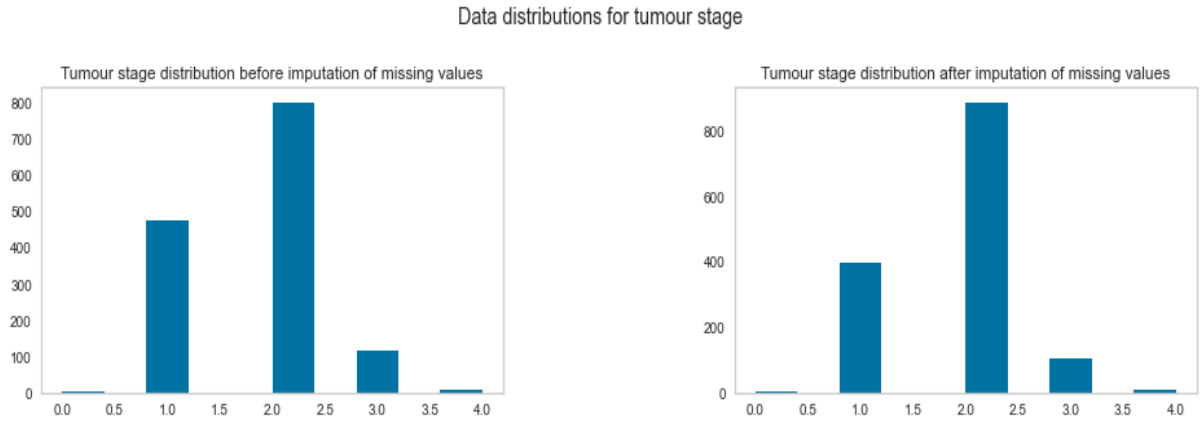
a:



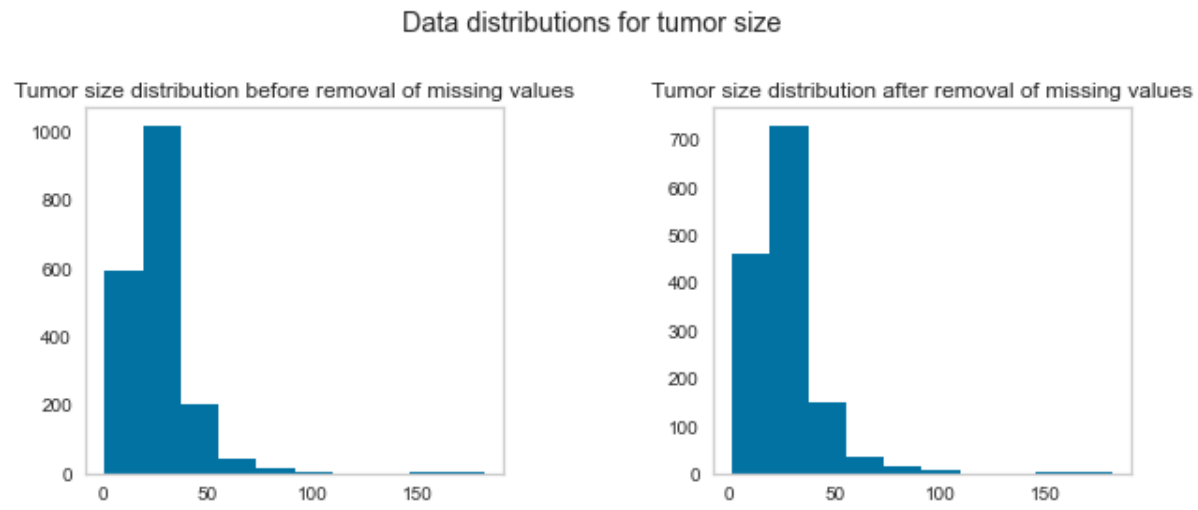
b:



c:



d:



A schematic indicating distributions of data before and after manipulation of missing values. Tumour size (fig. 2d) had 11 missing values, so the subjects with these missing values were dropped as little information would be lost from the dataset. The others were interpolated to retain the bulk of information from these subjects and will increase the predictive power of the models used.

Appendix D:

Table 2:

Machine Learning Model	Statistical dimensionality reduction employed	Root Mean Squared Error (RMSE) / months
Linear regression	None	104.9
Lasso regression	Incorporated in model	73.6
Ridge regression	None	66.8
Ridge regression	Principal component analysis	66.6
Polynomial regression	Principal component analysis (PCA)	216.2
XGBoost	Incorporated in model	55.8

A full description of models used for the prediction of breast cancer survival durations. The Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) database was employed, with XGBoost indicating the best performance.