**General Comments:**

**Internal Examiner Comments:**

**Comment 0.1 -** US/UK spelling, check modelled vs modeled, normalized v normalised - e.g. do a search for “ize” and replace with “ise” where applicable. Check Figure legends.

**Response 0.1 -** US spellings have been replaced by UK spellings. These changes can be seen in the PhD\_Thesis\_Diff file.

**Comment 0.2 -** Formatting chapter headings: “1 Chapter 1” not good.

**Response 0.2 -** I agree, in response I have changed the chapter heading format to “1 Introduction” etc. These changes can be seen in the PhD\_Thesis\_Diff file.

**Comment 0.3 -** Acronyms: define first then use forever. Lots of jumping back to long form, e.g. Kaplan-Meier then KM then Kaplan-Meier again.

**Response 0.3 -** In response I have made acronym use more consistent within thesis.

**Comment 0.4 -** Punctuation: overuse of commas instead of brackets or colons [page 54 great example].

**Response 0.4 -** In response I have used substantially more brackets (in place of commas) throughout the thesis.

**Comment 0.5 -** Acronyms in table and figure captions: need to be explained, this is mainly for Chapter 5 AI and AD models.

**Response 0.5 -** In response I have added acronym explanation in legends.

**Comment 0.6 -** Signposting: at the start (and maybe end) or each Chapter.

**Related to Comment 0.6 -** Signposting: better use of subsections to describe methods before results - think in terms of reproducibility.

**Response 0.6 -** I have added subsections to Chapter 2 and restructured chapter 3 to be easier to follow.

**Comment 0.7 -** Figure and Tables are usually numbered by Chapter rather than a running total - this might be helpful but not necessary.

**Response 0.7 -** I agree and have changed the numbering of Figures and Tables as such.

**External Examiner Comments:**

**Comment (Methods) -** Methods eg tests, visualisation, metrics reported, p-value adjustment etc Methods must be able to be replicated

**Comment (Equations) -** Equations must be numbered throughout eg pages 50/51

**Comment (Figures) -** Ensure legends, x/y axes labels on all figures. Ensure that multi-panel figures are labelled A and B etc (eg Figure 56)

**Comment (Legends) -** Legends for tables and figures must be detailed. Any abbreviations explained, datasets (or sections of) referenced.

**Comment (Sub-headings) -** Sub-headings would be useful and guide the reader through lengthy chapters.

**Response to above -** Have taken suggestions and altered thesis appropriately. See PhD\_Thesis\_Diff.

**Comment (Code) -** Code needs to be made available for all analyses, preferably in a code repository.

**Response (Code) -** Code has been made available on GitHub, link in Chapter 7 future work.

**Comment (Schemas) -** Schema/s would be useful to explain input/output and processing

**Response (Schemas) -** As we only use two pieces of software throughout chapters 5/6 and these pieces of software are used as directed in comprehensive documentation (including schemas), I have decided against including schemas in this thesis.

**Chapter 1: Introduction**

**Internal Examiner Comments:**

**Main Comment 1.1 -** You discuss the idea of improving predictive models - this doesn’t carry through later as we don’t compare prediction performance with/without CNA. Is this something for future work?

**Response 1.1 -** Yes, I agree that this is an aspect that could be looked at in the future. To make the reader aware of this I have added the following sentence to Chapter 7 Conclusions and Future Work”:

“Further to this, assessing the performance of our predictive models including the CNA Score and Burden metrics, molecular subtype classifications and selected clinical information as candidate predictors and comparison with predictive models only including the molecular subtype and clinical information, could be carried out.”.

**Main Comment 1.2 -** Clarify if Figures (e.g. Fig 1, 2) are original or from another publication in the caption.

**Response 1.2 -** To make it clear that these figures have been taken from other publications I have added “Taken from .....” to these figures, Figures 1 and 2.These changes can be seen in the PhD\_Thesis\_Diff file.

**Within Thesis Comment 1.3 -** Which predictive models? Are these from the literature or do you mean improve prediction models which only use clin and histo variables?

**Response 1.3 -** These predictive models are the ones including the clinical and gene expression data as alluded to in the thesis: “The aim of this thesis is to assess whether CNA information, in isolation or in combination with clinical and gene expression data, improves predictive models of overall survival (OS) and disease-specific survival (DSS) for breast cancer patients.”

**Within Thesis Comment 1.4 -** Why not give the exact number 2509?

**Response 1.4 -** There are 2,509 patients with clinical information available, but only 2,173 with CNA data available and 1,980 with gene expression data available. In addition, CEL files for allele-specific analysis were only available for 1992 patients. So, saying approximately 2,000 breast cancer cases covers the variation in availability of the different data types. In response no changes to thesis made.

**Note:** Also corrected typos/suggestions made in pdf of thesis.

**External Examiner Comments:**

**Main Comment 1.5 -** This chapter requires further extension and inclusion of the literature re molecular subtypes and omics signatures in BC.Consider molecular subtypes in greater depth. Also reference TCGA BC study (multi-omics).

**Response 1.5 -** In response I have expanded “1.1 Breast Cancer in the Clinical and Research Setting”, to include highlighted in red below:

In published research, gene expression and CNA data have been used to produce molecular classifications of breast cancer along with a number of prognostic and predictive assays, providing information about likely survival outcome and response to therapy \citep{pmid10963602, pmid22522925, pmid28882552}. Molecular-based classifications, being evaluated in the research setting, but not yet common place in routine clinical use, include the Prediction Analysis of Microarray 50 (PAM50) intrinsic subtypes and Integrative Clusters (IntClust) \citep{pmid10963602, pmid12829800, pmid22522925}. PAM50 is a 50-gene signature that classifies breast cancer into five molecular intrinsic subtypes, Luminal A, Luminal B, HER2-enriched, Basal-like and Normal-like, that have been shown to have both prognostic and predictive power. Briefly, using complementary DNA (cDNA) microarrays on 65 breast cancer samples \cite{pmid10963602} identified a subset of 496 genes whose variation in expression was significantly greater between samples from different tumours than between samples from the same tumour. Performing hierarchical clustering with this “intrinsic” gene subset resulted in the samples being split into four groupings related to different biological features (ER+/Luminal-like, Basal-like, HER2-enriched and Normal-like). Subsequently, using the same intrinsic gene set, \cite{pmid11553815} performed hierarchical clustering on 85 breast tumour cDNA microarrays. In addition to classifying samples into Luminal-like, Basal-like, HER2-enriched and Normal-like groups, the Luminal-like group was further divided into at least two subgroups, each with a distinctive gene expression profile. \cite{parker} later developed the 50-gene subtype predictor (PAM50) by performing gene set reduction on the 496 intrinsic genes \citep{pmid10963602, pmid11553815}, along with an additional 1,410 identified in three other microarray studies \citep{micro1, micro2, micro3}. Claudin-low, a sixth subtype of breast cancer identified using gene expression data in a separate study \citep{pmid17493263, pmid20813035}, is also considered an intrinsic subtype \citep{pmid32286297}. \cite{TCGA}, integrating DNA copy number arrays, DNA methylation, exome sequencing, messenger RNA arrays, microRNA sequencing and reverse-phase protein arrays, also observed the existence of four main breast cancer classes, each of which shows significant molecular heterogeneity in terms genetic and epigenetic alterations and are highly correlated with PAM50 subtype. IntClust derived from gene expression and CNA data classifies breast cancer into ten integrative clusters, IntClust 1-10, each with distinct CNA landscape, risk patterns and prognosis \citep{pmid22522925}. Prognostic biomarkers include ER, PR, HER2 and Ki67 status, Urokinase plasminogen activator/plasminogen activator inhibitor 1 (uPA/PAI-1), Oncotype DX, MammaPrint, Prosigna and Breast Cancer Index (BCI), while predictive biomarkers include ER status, PR status, HER2 status, deficiency in DNA damage response (DDR), mutational status of ER, amongst others \citep{pmid24402422, pmid28882552}.

**Main Comment 1.6 -** Consider predictive (response to treatment eg DDRD, Mulligan et al) versus prognostic (survival) models?

**Response 1.6 -** See above, response 1.5.

**Main Comment 1.7 -** Consider FDA/EMA-approved tests eg FoundationOne CDx (HER2-ve), GUardant360 CDx, PAM50 (Prosignia), DDRD/DDIR (Mulligan et al). Which tests are used in different clinics? Tests may be more widespread than you think. Specify which Omics levels these tests are based at i.e. transcriptomic or genomic.

**Response 1.7 -** While FoundationOne CDx and GUardant360 CDx can be used to predict response to therapy, they seem to be marketed as companion diagnostics applicable across a number of solid tumours, not just breast cancer. Furthermore, there does not seem to be adequate or accessible information on which tests are used in different clinics. As a result, the decision was made to not include FoundationOne CDx and GUardant360 CDx in the thesis.

**Within Thesis Comment 1.8 -** Why NAs? No gene expression, or could not be allocated PAM50/IntClust labels?

**Response 1.8 -** NAs occur when no gene expression data are available for patients. I have made the following highlighted changes to make this clearer in thesis Chapter 2:

“Within the METABRIC cohort there are 1,974 patients for which CNA data and PAM50 subtype information are available and 1,980 patients for which CNA data and IntClust information are available (Table 1). The 529 patients missing both PAM50 and IntClust information did not have any gene expression data available, meaning they could not be allocated PAM50 or IntClust, while 6 patients had IntClust information but were categorised as PAM50 “NC” and subsequently recoded as NA.”

**Within Thesis Comment 1.9 -** Table 3 - How is survival (median) calculated?

**Response 1.9 -** In Table 3 median survival time is calculated using the tbl\_summary() function and is just the median of all survival times regardless of censoring. As the survival time is treated the same as other continuous variables, I have not made changes to the thesis.

**Within Thesis Comment 1.10 -** Was the CNA/transcriptomic data re-processed? Or used as is? METABRIC training and/or test set? Were these combined? Batch effects of integrating two datasets directly?

**Response 1.10 -** The datasets were not reprocessed, and the test/training sets were combined, to make this clearer the following highlighted changes have been made to Section 1.2 of the thesis:

This publicly available data is highly curated and periodically updated with additional information or datasets. Throughout this thesis the processed METABRIC data downloaded from cBioPortal is used “as is” and with the training and test sets, defined in \cite{pmid22522925}, combined. Unless otherwise stated the results discussed in this thesis are based on the processed METABRIC data downloaded from cBioPortal in 2021.

**Chapter 2: Copy Number Alterations as a Measure of Genomic Instability**

**Internal Examiner Comments:**

**Comment 2.1 -** Need more detail on predictive accuracy of previous methods - how accurate are CES, CIN, CAAI for example? Can you provide simple accuracy/AUC statistics for the key methods?

**Related to 2.1 -** CES - how well - add %'s ?, give more detail on what you mean by "highly prognostic" - can you give %'s?

**Response 2.1 -** While it would be beneficial to provide these accuracy metrics for CES, CIN, CAAI etc., the original studies partitioned patients based on these genomic instability metrics and utilised KM curves, log-rank tests and Cox PH models to inspect the difference in survival between patients with varying levels of GI. They did not access predictive accuracy and so no accuracy/AUC statistics are available. As a result, this information cannot be added to the thesis.

**Comment 2.2 -** You say missing data are “unlikely to invoke bias” - what bias is being avoided? Clarify in text.

**Response 2.2 -** Discussed in viva. The bias that is being avoided here is the underestimation of the all-case CNA metrics for patients that have missing values in one or more genes. By calculating CNA metrics for these patients, the true value of the CNA metric is not being accurately described, i.e. CNAs may be missed leading to a lower score/burden than is actually present. However, for most CNA metrics the impact of missing values is minimal. To acknowledge this, I have altered the sentence (addition is in red): “it is determined that including all cases, across the global and chromosome arm-specific metrics, is unlikely to invoke bias in the form of underestimating the CNA metrics.”.

**Comment 2.3 -** Could missing data be imputed? Add a reason why this has not been done here.

**Response 2.3 -** While imputation could have been used, the proportion of missing data is small and the impact of the missing data on the majority of CNA metric distributions was shown to be minor. Therefore, imputation was not performed. To make this clear the line “it is determined that including all cases, across the global and chromosome arm-specific metrics, is unlikely to invoke bias in the form of underestimating the CNA metrics.” in the thesis was further changed to “it is determined that including all cases, across the global and chromosome arm-specific metrics, is unlikely to invoke bias in the form of underestimating the CNA metrics. Furthermore, imputation of missing CNA values was considered but not performed as the impact of the missing values on the CNA metric distributions was shown to be minor across the majority of CNA metrics.”

**Comment 2.4 -** If p>0.05 does this mean two groups are similar? Change to say “no significant difference”.

**Related to Comment 2.4 -** Using p<0.05 instead of reporting exact p-value is not recommended.

**Response 2.4 -** In response I have made relevant changes made throughout chapter 2. These changes can be seen in the PhD\_Thesis\_Diff file.

**Comment 2.5 -** What is new or novel in this chapter? Perhaps put a summary of this at the end.

**Response 2.5 -** I have added to conclusion “As the deletion landscape of breast cancer has been poorly characterised, our novel findings that PAM50 and IntClust classifications with poorer OS and DSS have higher levels of GI, in particular higher CNA deletion burden, encourage further investigation.”

**Within Thesis Comment 2.6 -** Probably don't need the decimals here?

**Response 2.6 -** I have kept these numbers to two decimal places to maintain consistency throughout thesis.

**Note:** Also corrected typos/suggestions made in pdf of thesis.

**External Examiner Comments:**

**Comment 2.7 -** Create table comparing GI measure/ metrics; suggested column headings: Cancer type Derived In/Output Files required, Signature/Test Name, Author.

**Response 2.7 -** In response I have included a table (Table 2.1) at the beginning of section 2.2 in Chapter 2. This table contains the following headings: GI measure, Cancer type(s), Input data, Platform used in study, Author.

**Comment 2.8 -** Why were these questions derived? Rationale? Why are these needed? Compare (more fully) to other metrics.

**Response 2.8 -** Discussed in viva. As mentioned in the thesis the existing GI measures "derived and evaluated using array-based data, aCGH and SNP arrays, or whole genome sequencing data, are limited in their accessibility and use as access to raw or segmented array/whole genome sequencing data is required. In addition, they can often be hard to interpret.”. These limitations of pre-existing GI measures encouraged us to create our own to hopefully provide researchers with easily interpretable and accessible measures of GI in cancer - “As a result, we aim to create easily interpretable GI measures that can be calculated using publicly available summary CNA data. These CNA metrics aim to capture the main aspects of CNAs, including magnitude, type and genomic location.”. Unfortunately, for many reasons it is difficult to compare our CNA metrics to the other metrics. The main reason comparison is difficult is that at the time the research was done, we did not have access to the raw or segmented data from which the other metrics are calculated.

**Comment 2.9 -** What does R overlap() function do? Fig 4/Table 9: concordance, define, how is it measured?

**Response 2.9 -** The overlap function, part of the overlapping package in R, returns the overlapped estimated area between two or more kernel density estimations from empirical data. The overlapping measure can be computed either as the integral of the minimum between two densities or as the proportion of overlapping area between two densities, i.e. where the integral of the minimum between two densities is divided by the integral of the maximum of the two densities. The overlap between the CNA metric distributions was first reported using the first method, the only one available before 2022 but has been updated subsequently to use the second method. See updated overlap metrics in thesis and addition of " This function is used to estimate the proportion of overlapping area between two densities, i.e. where the integral of the minimum between two densities is divided by the integral of the maximum of the two densities. This proportion is then multiplied by 100 to calculate the percentage overlap”.

**Comment 2.10 -** Tables 8,9: made it clear that you are talking about full vs CC in legends.

**Response 2.10 -** In response I have changed legends to “The percentage overlap between the global complete-case and all-case CNA Burden metric densities. Metrics are ordered and coloured by percentage overlap.”

**Comment 2.11 -** “Rather than extensively presenting details of distributions for each of the 42 chromosome arms, we select chromosome arm 1q for more detailed illustration and discussion.” Rest should go into supplementary, can be online.

**Response 2.11 -** In response, I provided a link to online supplementary information including the 41 other chromosome arm figures.

**Comment 2.12 -** This is telling you that your CNA metrics are associated with PAM50/IntClust. Are the CNA metrics 'better' at explaining survival outcomes than PAM50/IntClust?

**Response 2.12 -** While the CNA metrics are associated with PAM50 /IntClust, it is unlikely that the CNA metrics are better at explaining survival outcomes PAM50/IntClust. There are several reasons for this, primarily that the CNA metrics only quantify the levels of CNAs across the genome and give no consideration to the biological impact of these CNAs, i.e. gene expression changes. In contrast the PAM50/IntClust includes gene expression data which provides vital information on the biology of the tumour which results in the PAM50/IntClust being better able to explain survival outcomes. Furthermore, large variation exists in CNA metrics observed across and within PAM50/IntClust.

**Comment 2.13 -** P36 Did you apply K-W, adjust and present those that reached significance? How did you adjust? B&H? Did you adjust for multiple K-W tests and then present adjusted p-values for the post-hoc tests?

**Response 2.13 -** Both the K-W and Dunn’s test p-values were adjusted using B&H. I have changed the thesis in multiple places to make this clearer (see latex diff file).

**Comment 2.14 -** P37 Survival 'The remaining Int-Clust classifications with favourable survival outcomes i.e. IntClust 3, 4ER-, 7 and 8 have similar levels of amplifications and deletions, IntClust 3, 4ER-, 7, p > 0.05, or significantly more amplifications than deletions, IntClust 8, p < 0.0001' Was this your assessment of survival outcomes or original authors'?

**Response 2.14 -** This assessment of survival outcomes was from the METABRIC publication. I have updated thesis to recognise this: “The remaining IntClust classifications observed by \cite{pmid22522925} to have favourable survival outcomes, i.e. IntClust 3, 4ER-, 7 and 8, show no significant difference in the levels of amplifications and deletions, IntClust 3, 4ER-, 7 ($p > 0.05$) or significantly more amplifications than deletions, IntClust 8 ($p < 0.0001$).”

**Comment 2.15 -** Figure 18 needs to go earlier in chapter when introducing PAM50/IntClust

**Response 2.15 -** In response Figure has been moved to chapter 1.

**Comment 2.16 -** Pg 40: “some noteworthy alterations primarily observed in Basal patients include high levels of amplifications on chromosome 3q”. Anything else can go in supplementary.

**Response 2.16 -** Figures focusing on noteworthy alterations observed in the different PAM50/IntClust classifications were specifically produced for chromosome arms displaying subtype specific alterations. These subtype specific alterations were identified by looping through chromosome arm and CNA metric combinations and producing density plots. As a result, no other subtype specific figures, like Figure 21, were produced, and including all chromosome arm and CNA metric combination density plots in the supplementary information would make the supplementary information exceptionally large for very little gain/additional information.

**Comment 2.17 -** Figure 21, Basal-focus, but all subtypes in plots. Need to alter to bring greater focus to Basal.

**Response 2.17 -** While Figure 21 (Now figure 2.20) is focused on the Basal patients, I think including the other PAM50 subtypes in the density plot is beneficial to get an idea of the levels of GI in Basal patients compared to other subtypes. In response to this comment though I have altered the legend to make it clearer that the focus is on Basal patients. Changed to “Density plots for each selected chromosome arm CNA Burden metrics, with a focus on the Basal subtype, accompanied by adjusted Kruskal-Wallis p-values. Chromosome arms where Basal patients display high GI are selected. Each facet contains boxplots for the chromosome arm CNA Burden metrics calculated using all available data.”

**Comment 2.18 -** Table 20, How is this a focus on the Basal type, all pairs are presented? Need to highlight in legend that you are picking out c arms in which Basal pts demonstrated high GI - highlight basal pairs in bold. The non-basal pairs are presented as a comparison.

**Response 2.18 -** In response have highlighted row corresponding to the PAM50 subtype of focus in each of the tables, Tables 2.18, 2.19 and 2.20. I have also altered the legend to put more focus on subtype of interest. Changes as follows - “Comparisons of selected chromosome arm CNA Burden metric distributions by PAM50 subtype, with a focus on the Basal subtype. Chromosome arms where Basal patients display high GI are selected. Z statistics and adjusted p-value, adjusted within each metric, for each Dunn's test are shown.”

**Chapter 3: Association of Copy Number Alteration Signatures and Survival Outcomes**

**Internal Examiner Comments:**

**Comment 3.1 -** What are the measures of node impurity for categorical data - Gini info and entropy - can you briefly explain.

**Response 3.1 -** In response I have provided more information in the thesis by addition of - “For categorical response variables measures of impurity may be the Gini index or information index. The Gini index denotes the probability that a random observation is misclassified when chosen randomly, while the information index relates to how much information is gained by splitting a set of data on a particular feature.”.

**Comment 3.2 -** Use past tense in describing results. Analysis is done in steps and present tense does not work well.

**Response 3.2 -** I changed the present tense that was highlighted in the text for Chapter 3. I have made no other changes as I was not sure if this comment was specific to Chapter 3, or changes were meant to be made across the entire thesis. If further changes are necessary, please let me know.

**Comment 3.3 -** Provide HR and CI from Cox models rather than just p-values.

**Response 3.3 -** In response, I added HR and CI to several places in the text (within the Luminal breast cancer section (3.2)). See PhD\_Thesis\_Diff file.

**Comment 3.4 -** What is the depth of a tree (page 66 depth of 4)?

**Response 3.4 -** The depth of the tree corresponds to how many levels/layers it has. To make this clearer I have added “To avoid overcrowding, these trees are limited to a depth of four, where depth is defined as the number of layers or levels in the tree.”.

**Comment 3.5 -** Talk through model building in Section 3.2 … univariate models used to find significant variabless, then dropped “similar variables”, then CNA score still significant. But why drop anything if prediction is the goal, and why use significance if inference is the goal? Needs more detail, perhaps in a subsection?

**Response 3.5 -** In this section we wanted to see if Absolute CNA Score was associated with survival in Luminal patients. As a result of this we wanted to minimise collinearity in the variables used in the Cox models and so association tests, prior knowledge and clinical information was used for variable selection. More information about this can be found in King et al., 2021. In response I have made this clear by addition of “discussed further \cite{King\_2021}” in relevant places in thesis.

**Within Thesis Comment 3.6 -** What are the consequences of violating the Cox PH assumptions?

**Related Within Thesis Comment 3.6 -** What's the consequence of violating the PH assumption? Is it ever valid? There's literature on this question that might be of interest. In essence I think it reduces your statistical power to detect an effect. Since you find a significant effect, I'm 0.not sure it's a big deal here!

**Response 3.6 -** In response have added the following sentence to thesis: “Violations of these assumptions may lead to inaccurate results, i.e. biased effect estimates, and reduced statistical power \citep{pmid35070171}”

**Note:** Also corrected typos/suggestions made in pdf of thesis.

**External Examiner Comments:**

**Comment 3.7 -** More references with respect to definitions/explanations. Parametric techniques are introduced, but no examples provided.

**Response 3.7 -** In response I have added more references to the survival analysis methods section at the beginning of chapter 3. Examples of parametric techniques were also added. These changes are highlighted as additions in PhD\_Thesis\_Diff (latex difference file).

**Comment 3.8 -** P50 What are commonly used/recognised predictors (prognostic) with respect to BC? (Or subtypes).

**Response 3.8 -** The sentence referenced was meant to highlight the clinical variables used as predictor variables in survival analysis in breast cancer. To make this clears I have changed the sentence to “Some examples of clinical predictor variables used in the context of breast cancer survival include number of lymph-nodes positive, tumour size, tumour grade, tumour subtype, ER/PR/HER2 status and type of treatment.”. This change is highlighted as an addition in PhD\_Thesis\_Diff (latex difference file).

**Comment 3.9 -** P51 Did you test assumptions with respect to the METABRIC dataset? If visual or stats tests, include them in supplementary.

**Related Comment 3.9 -** P56 'Assuming the assumption of proportional hazards, the results indicate that the CNA Score metric is significant in a model for DSS along with six clinical predictors, PAM50 subtype, histological grade, tumour size, number of positive lymph nodes, age at diagnosis, and HER2 status, both using the categorical CNA Quartiles, Table 28, and the original CNA continuous variable' Were assumptions tested?

**Related to Comment 3.9 -** P58 'diagnostic tests indicate that the proportional hazards assumption may not be met.' Plots need to be in Supplementary, results from any Cox PH analyses

**Response 3.9 -** We did test the Cox PH assumption for the METABRIC Luminal breast cancer patients. The assumption was tested using the Scaled Schoenfeld residuals, tables now provided in the online supplementary materials (Table 2.1 and Table 2.2).

**Comment 3.10 -** P52 Recursive partitioning Survival Trees, how are decisions made to split at a node, via log-rank test?' Or response = 0,1? If latter, is this appropriate, or could OS60 or RFS60 be used?

**Response 3.10 -** The rpart recursive partitioning survival tree uses impurity measures such as the Gini index or information impurity to choose the partitioning variable and optimal split giving the lowest amount of impurity. In comparison, the ctree recursive partitioning survival trees use a significance test procedure to select variables to partition on. Both these trees support survival data (time and censor status) and throughout the thesis we use overall, 5- and 10-year OS and DSS survival data.

**Comment 3.11 -** P54 Quartiles, why? By applying thresholds to continuous data, there is a risk of loss of information. Issues at boundary points too. Are quartiles appropriate given distribution? If randomly-split data into four equally-sized groups and repeated multiple times (n = 100) how many times would significance be reached?

**Response 3.11 -** Discussed in viva. Briefly, cases are assigned to ranked quartiles as a first order means of segmentation for downstream analysis making comparison between patients with varying levels of GI easier to interpret and visualise. As this results in information loss and may not be completely appropriate given distribution, results of considering the continuous Absolute CNA Score are presented alongside the Absolute CNA Score Quartile results. In addition, in subsequent analysis sections only the continuous CNA metrics are utilised. No amendments to thesis made.

**Comment 3.12 -** P56 'Since highly correlated predictors may lead to unreliable and unstable estimates of regression coefficients, a refined selection of variables are considered, based on understanding of the clinical definition of the variable, e.g. HER2 Status and HER2 SNP6 use different methods to capture similar information.' Need to consider not only associations with CNA score, but also associations between clinico-path variables.

**Response 3.12 -** Discussed in viva. While this is an important consideration and something to be aware of going forward, apart from additional information detailing variable selection provided in King et al., 2021, no amendments to the thesis were made.

**Comment 3.13 -** 'For Luminal B patients, the effect of CNA Score on DSS is estimated by fitting interaction effects between CNA Score and PAM50 subtype.' Why and how? Already shown that association between Luminal cat + CNA scores

**Response 3.13 -** Discussed in viva. Based on ambiguity in luminal classification, we wanted to see if the CNA Score metrics had different effects on Luminal A versus Luminal B patients. Interaction effect was added to explore this.

**Comment 3.14 -** If Pearson's used, were assumptions tested? It was shown that PAM50 was associated with CNA score, but this is different from an interaction. Did you test the interaction in a Cox model, prior to introducing other variables i.e. ~ A+B+ A:B?

**Comment 3.14 -** Discussed in viva. When using Pearson’s correlation test, we did not test assumptions. This will be implemented going forward/in future work. Yes, we did the interaction in a Cox model, prior to introducing other variables.

**Comment 3.15 -** Was the concordance index evaluated for Cox PH models? Did you compare non-CNA MV models with CNA MV models?

**Response 3.15 -** Discussed in viva, yes, we did compare concordance index and saw small improvements with addition of CNA Score/Quartiles.

**Comment 3.16 -** Surv Trees: was there a random effect associated with the set up of each tree eg set.seed(). If so, were there major deviations if different seeds used.

**Response 3.16 -** Discussed in viva. We used set.seed for the ctree algorithm (rpart algorithm does not require set.seed). Changing seed does not lead to major changes in tree structures.

**Comment 3.17 -** Given that IntClust was developed using both CNA and GE, what does the split by CNA event from INTCLUST factor signify? Were the Curtis groups too rigid? Are there issues with respect to using two variables, one of which is a contributor to the other?

**Response 3.17 -** Discussed in viva. Using the CNA metrics and IntClust together help identify cases where there are sub cohorts of patients within the cluster, that have different survival profiles. The survival trees are not affected by high multicollinearity amongst predictors and so there are no issues with including the two variables in the trees.

**Comment 3.18 -** Five-year data - explain how this was achieved eg if patient experienced an event at 61 months, then they would be censored (0) at 60 months. Adaptations to data must be detailed in methods.

**Response 3.18 -** Discussed in viva.To make readers aware of how the 5 and 10-year OS and DSS was calculated I have added the following sentence to the introduction chapter in the thesis: “From the OS and DSS variables, 5- and 10-year OS and DSS are generated for each patient, e.g. if a patient experienced an event at 5 years and 1 month (61 months), then they would be censored (0) at 5 years (60 months).”

**Comment 3.19 -** Who is this aimed at? Also stepAIC() or lasso for var selection.

**Response 3.19 -** GNOSIS is aimed at researchers in either a clinical or academic setting. This is mentioned in thesis “developed to enable the exploration, analysis and incorporation of a diverse range of genomic features with clinical data in a research or clinical setting”. Currently GNOSIS does not have the ability to perform StepAIC() and LASSO, this will be considered in future updates.

**Chapter 4: Effect of Copy Number Alterations on Gene Expression**

**Internal Examiner Comments:**

**Comment 4.1 -** How were the trees for DGEA chosen (end of page 123)? Clarify.

**Response 4.1 -** A range of trees across survival times (DSS, 5- and 10-year DSS) and algorithms (rpart and ctree) were selected for DGEA. The trees discussed in the thesis were selected because they partitioned the patients once on PAM50 or IntClust and subsequently on one global or per arm CNA metric. As a result, DGEA is carried out comparing nodes that consist of low GI PAM50 subtypes or IntClust that were further partitioned using a CNA metric. This makes it easier to analyse the effect of one specific CNA metric on subsets of patients, rather than comparing nodes arrived at through partitioning using multiple classifications or CNA metrics. To make this clear the sentence “These survival trees, produced for a range of survival times (DSS, 5- and 10-year DSS) and algorithms (rpart and ctree), were selected for DGEA as they partitioned the METABRIC patients first on PAM50 or IntClust and subsequently on a single global or chromosome arm CNA metric, simplifying comparison and inference.” is added to thesis.

**Comment 4.2 -** Is this model run separately for each gene? Clarify.

**Response 4.2 -** A linear model is run for each gene in both limma and our modified version of limma. The only difference is that in our modified version each linear model has a different design matrix. It is mentioned in section 4.1 that in limma “The linear model is then fitted for each gene given a series of arrays, with the option to apply array weights.” and this holds for our modified limma.

**Comment 4.3 -** Discuss comparative results - should more/less overlap be expected?

**Response 4.3 -** Comparative results are discussed in thesis on pg. 131/133. In brief modest overlap is expected for the prognostic and predictive assays due to the difference in objectives of each study. Of course, some overlap will occur as oncogenes and tumour suppressor genes (which are likely to be included in prognostic and predictive assays) are also likely to be impacted by CNAs, where the CNA influences gene expression. On the other hand, high levels of congruence between the IntClust gene set and our gene set was expected for several reasons, mainly that the IntClust gene set was also produced using the METABRIC CNA and gene expression data and although the authors used a different approach, the idea was similar.

**Comment 4.4 -** Can gene expression be used to improve predictive ability? Future work?

**Response 4.4 -** I agree that including gene expression in our predicative models. See chapter 7 comments.

**Comment 4.5 -** Is it worth pursuing the five-state specification then?

**Response 4.5 -** A large number of genes have small sample size in the five-state specification. However, as our differentially expressed gene set is determined as “differentially expressed genes of sufficiently large sample size, i.e. genes with an adjusted p-value < 0.05 and absolute log fold change > 0.58 in at least one contrast” it allows genes who are up or down-regulated to be captured in any of the comparisons. This means that if the gene is removed in one contrast it can still be detected in the others. Indeed, an extra 2,179 genes are captured as differentially expressed in the five-state specification compared to the three-state specification. Fo this reason it seems like pursuing the five-state would be informative.

**External Examiner Comments:**

**Comment 4.6 -** PAM50 etc was derived from clustering rather than DEG.

**Response 4.6 -** In response I have amended thesis, i.e. removed mention of PAM50 being derived from DGEA. See PhD\_Thesis\_Diff file.

**Comment 4.7 -** DE gene lists need to be presented in supplementary, with p-values/adjusted p-values/q-values.

**Response 4.7 -** The ModLim3 and ModLim5DE gene lists are now provided in the supplementary information in the form of .txt files.

**Comment 4.8 -** Include up/down-regulated. Includes numbers of up/down-regulated in main text, rather than 'a number'.

**Response 4.8 -** I have changed “reveals a number of differentially expressed genes” to “reveals \*\* down-regulated and \*\* up-regulated genes.” throughout Chapter 4. See PhD\_Thesis\_Diff file for changes.

**Comment 4.9 -** Why LIMMA? Why not SAMR?

**Response 4.9 -** Discussed in viva. In short, as the METABRIC study itself used limma and I had significantly more experience using limma than SAMR, the decision was made to use limma. In addition, as the documentation for the limma package is extensive and quality of code high, I felt more comfortable altering the code base of limma to produce our modified limma.

**Comment 4.10 -** What was the rationale for choosing the trees (eg 3p).? 'We first focus on presenting four DGEA applications of interest from the survival trees produced as a result of incorporation of the global CNA metric information and then on presenting three DGEA applications of interest from the survival trees produced as a result of incorporation of the chromosome arm CNA metric information.'

P123 State how many trees that you are going to look at with respect to DGE.

**Response 4.10 -** For the chromosome arm survival trees, 3 survival trees were selected, for chr 3p, 18q and 11p. The rationale for choosing the trees (eg 3p) is similar to internal examiner comment 4.1. In short, we wanted to compare nodes created using relatively simple partitioning, i.e. on molecular classification and then one only one CNA metric. In addition, these survival tree arms consistently appeared as useful in partitioning patients based on survival. In response altered thesis to include number three in “DGEA is applied to three survival trees of particular interest.”

**Comment 4.11 -** P126 Define beta, epsilon

**Response 4.11 -** In response I have amended thesis by adding “where $\textbf{X}$ is the design matrix and is the same for each gene $g$, $\beta$ is the vector of parameters and $\epsilon$ is the error vector.”. See PhD\_Thesis\_Diff file.

**Comment 4.12 -** Clarify, are multiple LIMMAs being carried out eg group A = gene1 amplified, group B = gene1 neutral, then group C = gene 1 deleted, group B = gene1 neutral, with Group A vs B and Group C vs B, and repeat for multiple genes/events.

**Response 4.12 -** See internal comment 4.2. In short, limma runs an individual linear model for each gene with the same design matrix, our modified limma works the same way, i.e. runs a linear model for each gene, but the design matrix for each gene is different. All comparisons are carried out at once and the comparison of interest is specified using contrasts.

**Comment 4.13 -** How are batch effects dealt with re GE in METABRIC datasets?

**Response 4.13 -** Discussed in viva. Batch effects have not been dealt with but that is a potential consideration in future work.

**Comment 4.14 -** Why choose this threshold?

**Response 4.14 -** Discussed in viva, this seems to be the threshold used in limma tutorials and literature. I have added references to thesis to back this up, see PhD\_Thesis\_Diff file.

**Comment 4.15 -** P129/130 Could there be surrogates, or downstream effects, of events at genome level. That is, could genes in other signatures be correlated with genes highlighted in thesis?

GE data will incorporate signals from genome, methylome and transcriptome levels. Genome only one part.

Could these genes outperform CNA with respect to survival association? What else could you do with these genes? Are they correlated with each other?

Could they be used in clustering (similar to intClust?)

Functional enrichment analysis? Shared biology?

**Response 4.15 -** Discussed in viva. Genes highlighted in this thesis could very well appear in other signatures. These genes are genes where the presence of a CNA may influence gene expression. As noted in Curtis et al., 2012, these genes will be enriched for oncogenes and tumour suppressor genes which are commonly included in prognostic/predictive signatures.

Although we haven't tested it, these genes could outperform our CNA metrics. This is because as mentioned in comment, the gene expression level does incorporate multiple layers of information (genome, methylome, transcriptome).

While we could perform clustering with these genes, as our gene set were highly correlated with the IntClust gene set (as expected), similar clusters would likely be produced.

**Chapter 5: Modelling Allele-Specific Copy Number Associated Changepoints**

**Internal Examiner Comments:**

**Comment 5.1 -** Why curtail number of CNA to amp, neut, del? Changepoint models could be applied to CNA counts. Future work?

**Response 5.1 -** Discussed in viva. The decision was made to curtail the copy number changepoints to deletion, amplification and neutral to focus on cases where the copy number changed state completely, i.e. deletion to amplification, as opposed to amplification to amplification of a higher magnitude. We acknowledge that by curtailing the CNA state to amplification, deletion or neutral, changepoints corresponding to amplification to higher/lower magnitude amplification are missed. As these changepoints could hold valuable information and should be investigated, the following sentence was added to future work: “In addition, during preprocessing of the ASCAT data the copy numbers of each allele were bound in the range [0-2], resulting in copy number changes for amplified regions being missed. These changepoints could potentially provide valuable insights and their impact on survival and treatment response should be investigated in future work.”

**Comment 5.2 -** AINM models do not use a random intercept for individuals, has this been tested? Discuss.

**Response 5.2 -** Themodels including intercepts do not use a random intercept for individuals. The reason for this was mainly down to time constraints. However, the use of random intercepts and the use of random effects in general could be considered in future work.

**Comment 5.3 -** Need to walk through the simulation study - are the estimates correct - what is the truth? 0=0 is fine, but the others with values are highly variable … is the mean where it’s supposed to be?

**Response 5.3 -** I assume this is related to the dot plots (Figure 5.12 and Figure 5.13). In short, the mean is where it’s supposed to be. These means were generated using the truncated Normal distribution with a large standard deviation. In the simulated example N = 50 and P = 20, resulting in 10 non-zero observations in each dataset. As the sample size is small the mean is highly variable. However, it is being predicted correctly and that's what we are interested in. If I have misunderstood comment, please let me know.

**External Examiner Comments:**

**Comment 5.4 -** Explain the 'why' of Chapters 5/6 - how can the info generated be taken forward?

**Response 5.4 -** In response have added to sentence in introduction to chapter: “In this chapter, measurements of allele-specific copy number profiling for the METABRIC cohort, using Allele-Specific Copy number Analysis of Tumours (ASCAT) \citep{pmid20837533}, is presented and a modelling framework is proposed for the detection and classification of changepoints in observed allele-specific copy number profiles, with a simulation study carried out to assess modelling approaches. These analyses aim to capture large scale changes in CNA state that occur on individual alleles, providing information on CNA events that may occur preferentially in certain genomic regions and whose downstream influence requires more investigation.”

**Comment 5.5 -** P133/134 'As discussed in section 2.2.8, Pladsen et al. (2020) utilised allele-specific copy number profiles, generated using ASCAT, to produce six metrics, AMP, DEL, STP, CRV , LOH, and ASM. The AMP, DEL, STP and CRV metrics were calculated on the sum of the allele-specific copy number, and all six metrics were combined into two prognostic indices, CPI and CPIweighted. n, combining the metrics into a single index result in loss of valuable information, specifically the type of CNA observed and which allele the CNA is observed on.'

**QUESTION:** Compare with quartiles applied in thesis CNA metrics/scores.

**Response 5.5 -** Yes, although it was not done, it would be possible to compare the CNA Quartiles to the tertiled CPI and CPI\_weighted. This would indicate whether our gene-centric total CNA Quartiles contain as much information as the tertilesd CPI metrics generated using whole genome allele-specifc CNAs.

**Comment 5.6 -** PennCNV/ASCAT Broad picture: why chosen and how do they work?

**Response 5.6 -** Discussed in viva. ASCAT was chosen for a number of reasons including that it is applicable to Affymetrix SNP6 data without matched normals, it is well documented and used widely in publications and it provides outputs, such as the segmented copy number file, that allow us to extract the changepoint information easily. The PennCNV software was chosen to generate the LRR and BAF files (that are used as input to ASCAT) primarily because the creators of ASCAT use PennCNV in their pipeline and recommend it’s use, particularly for the microarray pipelines. In the thesis I have mentioned this information in “Giving consideration to software packages available for analysis of microarray data, specifically Affymetrix SNP6 data without matched normals, the type of output, and quality of documentation, ASCAT was deemed as most suitable for allele-specific copy number calling in this study.” and “When preprocessing the Affymetrix SNP 6.0 CEL files, for input into ASCAT, the PennCNV pipeline recommended by the creators of ASCAT, and subsequently applied here, contains three substeps, substeps 1.1, 1.2 and 1.4.”.

**Comment 5.7 -** P136, Figure 91 Explain purity% + goodness of fit%

**Response 5.7 -** I have amended figure legend to include “Ploidy is defined as the amount of DNA relative to a haploid genome and purity is defined as the percentage of tumour cells. These ploidy and purity estimates are generated by creating a grid of possible values and evaluating the goodness-of-fit for both parameters, for more detailed information see \citep{pmid20837533}.”

**Comment 5.8 -** P145 Expand on limitations of predict() function with respect to lm()

**Response 5.8 -** Discussed in viva and I’m not sure there is much to expand on. Obtaining prediction intervals for multivariate models fitted with the lm() function is not supported, i.e. only prediction point estimates are produced, meaning that we can’t use interval approach for significance. I have not made any changes to thesis as this is mentioned multiple times in thesis:

“Obtaining prediction intervals for multivariate models fitted with the lm() function is not supported.”, “Notably, the lm() and MCMCglmm() functions perform similarly across the univariate AI models, but limitations in the lm() and predict() functions result in only the MCMCglmm() function being capable of producing prediction intervals for multivariate models.” and “The multivariate AD models, fitted using the lm() function, were lacking prediction intervals, a result of limitations in the software.”

**Comment 5.9 -** P147 Explain interaction terms with respect to AMinor.

**Response 5.9 -** I think I have already done that in thesis: “where the term AlleleMinorij , corresponds to an indicator term with value 1 if the observed changepoint ij comes from the minor allele, and the estimated coefficient β7 corresponds to the estimated difference in response length for the minor allele compared to the major allele, within the NoChangepoint category, β0 .”

**Comment 5.10 -** Define Fit? Careful of abbreviation eg LB/UB - need to be in table legend Legends need to be standalone from main text. Explain all column headers in legend (applies to all tables).

**Response 5.10 -** In response I have defined fit in the thesis by adding to relevant table legends: “Fit, LB and UB correspond to the parameter estimate and associated 95\% interval.”.

**Chapter 6: Application of Allele-specific models to the METABRIC data**

**Internal Examiner Comments:**

**Comment 6.1 -** KM analysis: are changepoints good or bad? Conflicting results in presentation. Discuss.

**Response 6.1 -** Discussed in viva. Preliminary analysis suggests the presence of specific changepoint categories in genes may be associated with decreased DSS, while the presence of specific changepoint categories in genomic segments may be associated with improved DSS. There are several possible explanations for these conflicting results:

* These preliminary KM curves only consider one type of changepoint at a time, i.e. compares survival of patients who have specific changepoint with patients who don’t. However, there could be other changepoints occurring that are influencing survival that are not being captured.
* When whole genome/segment duplication or deletion happens no changepoint events occur. As a result, these genomic events are not captured. This is related to point 1 in the sense that instability may be occurring, but it is not being captured by the KM curves.
* The KM curves do not include/consider any other variables such as clinical variables.

Basically, we are only getting a very limited view of what's going on and so more investigation/research should be carried out. This is discussed in the thesis:

“However, there are a number of explanations for this: the KM curves only consider one type of changepoint at a time, i.e. compares survival of patients who have specific CNA changepoint with patients who do not, even though there could be other changepoints events occurring in a genomic region simultaneously, the KM curves do not consider any other variables such as clinical variables, and whole genome duplication events or changepoints with large lengths occurring in adjacent segments (resulting in CNAs spanning the length of the segment ) are not detected.”

**Comment 6.2 -** Could changepoints and CNA Scores be used in combination to predict survival? Discuss.

**Response 6.2 -** I’m not sure how you could combine the two into one model? CNA Score is calculated for each patient while the CNA changepoints are detected for each patient but then combined to detect regions with changepoint of significant length. I suppose you could count the number of changepoints of significant length for each patient and combine this information with CNA Score/Burden.

**Within Thesis Comment 6.3 -** why this tree? why not use the clinical variables?

**Comment 6.3 -** These trees are the same trees we selected in Chapter 4. Rational behind selecting these trees, where chromosome arm CNA metrics were used as candidate predictors, include ease of interpretation as they partitioned the patients once on PAM50 or IntClust and subsequently on only one chromosome arm CNA metric. In addition, chromosomes 3p and 18q consistently appeared as useful predictors of survival.

**Minor Comment 6.4 -** Frequency of changepoints would be a better label? The y-axis is different for each CP type and total, which makes it confusing. Could you use the same scale for all?

**Response 6.4 -** Agree, in response I have changed label to “Frequency of Changepoints”. In addition, I have used the same scale for all facets.

**Within Thesis Comment 6.5 -** Could you add the survival functions to the top of the plot?

**Response 6.5 -** IAgree, in response I have added the survival tree plot on top of the node plot.

**Minor Comment 6.6 -** it's very difficult to take anything from this graph as there's a lot going on and the colours on the bars are impossible to see. Could this be better as a table?

**Response 6.6 -** I agree it’s difficult to see the colours relating to the different categories of changepoints. However, the aim of these histograms is to show the frequency of changepoints across the genome and highlight regions where the frequency of changepoints is increased. As we are interested in the distribution of changepoints across the genome, using a table, which would end up being very large, is not appropriate. In response I have tried to make the histogram clearer by removing the colours/legend and making the figure bigger.

**Minor Comment 6.7 -** Minor allele - what does this tell us? is this useful?

**Response 6.7 -** This isn’t that useful, the fact that more deletions were observed on the minor allele is expected as the minor allele is defined as the allele with the lowest average copy number across the whole genome.

**Minor Comment 6.8 -** LCE1E - what does this tell us - is this linked to BCa?

**Response 6.8 -** LCE1E is not linked to breast cancer and many genes identified as having changepoints of significant length in the thesis have not been associated with breast cancer. The fact that they have a changepoint does not automatically mean there are downstream effects etc. Further investigation into these genes and their associations with survival in breast cancer is warranted. I have added “These changepoints could potentially provide valuable insights and their impact on survival and treatment response should be investigated in future work.”

**Minor Comment 6.9 -** Does this stand up to multiple testing? no adjustment for other parameters.

**Response Minor Comment 6.9 -** The p-value shown in thesis is unadjusted and does not stand up to multiple testing. In response I have added “However, when multiple testing correction is applied, the p-value becomes non-significant at α = 0.05” to thesis. SeePhD\_Thesis\_Diff file.

**External Examiner Comments:**

**Comment 6.10 -** P107 Heatmap based on clustering - distance/linkage used? Is it possible to visualise based on PAM50/IntClust subtypes?

**Response 6.10 -** The heatmap is based on clustering performed using the ComplexHeatmap package. The distance and clustering methods used were the defaults ("euclidean" and "complete"). The PAM50 and IntClust composition of the heatmaps are provided in previous heatmaps and is the same. I will add it for consistency.

**Comment 6.11 -** P167 'We first test whether the mean lengths of the changepoint alterations, TS and TE, are greater than 10kb, where significance is determined by the lower bound of the prediction interval for the relevant length, TS or TE, being greater than 10kb, LB > 10kb' Explain test in greater detail?

**Comment 6.11 -** The test is described in Chapter 5 and is related to our models. Basically, once the model is fitted, if the estimated parameter, i.e. the beta (which represents the mean length of the changepoint category) is greater than some specified length then the changepoint is called as significant.

**Chapter 7: Conclusions and Future Work**

**Internal Examiner Comments:**

**Comment 7.1 -** Can you use GE + CNA score for prediction?

**Response 7.1 -** Yes, you could use both CNA Score and Gene expression in combination. In Response I have added the following sentence to thesis: “The gene expression analyses carried out using limma (the survival tree node analysis and CNA state analysis) were performed including only gene expression and CNA state in our models. Future work may include application of other gene expression methodologies, e.g. Significance Analysis of Microarrays, expansion to consider other variables of interest, and the possibility of combining gene expression data and our CNA metrics in predictive models for survival outcome. ”

**Comment 7.2 -** Can you use combine parts 1 + 2 for prediction?

**Response 7.2 -** See response 6.2.

**Comment 7.3 -** What about allowing copy number > 2 in changepoint detection?

**Response 7.3 -** Yes, I agree this should be considered going forward. As a result, I have included the following sentence in the Future Work section: “In addition, during preprocessing of the ASCAT data the copy numbers of each allele were bound in the range [0-2], resulting in copy number changes for amplified regions being missed. These changepoints could potentially provide valuable insights and should be included in future work.”

**External Examiner Comments:**

**Comment 7.4 -** This chapter needs to be extended. Considering the amount of work that has been produced, it would be beneficial to see how the results stand with respect to the literature.

**Response 7.4 -** In response I have expanded the chapter by including information requested in subsequent comments. See below.

**Comment 7.5 -** Further discussion of benefits/limitations of thesis-defined CNA metrics versus those currently available.

**Response 7.5 -** I agree, in response have added the following to Conclusions:“While it was shown that our CNA Score and Burden metrics are accessible, easy to interpret and can provide valuable insights about how GI can impact survival outcomes, when comparing to pre-existing CNA measures, a number of limitations exist. These limitations, not encountered in most of the CNA measures discussed in Chapter 2, include that our CNA metrics are calculated using CNA summary data for annotated genes only, meaning the length of the CNA is not considered and CNAs occurring outside of gene regions are not included.”. See PhD\_Thesis\_Diff files.

**Comment 7.6 -** Cover both theoretical and application advantages/disadvantages.

**Response 7.6 -** Limitations of our selected model (ADIM) are discussed in future work section.

**Comment 7.7** - What other papers have looked at novel CNA metrics? (Not just in BC).

**Response 7.7 -** All of the pre-existing measures of genomic instability, i.e. pre-existing CNA metrics, mentioned in the thesis have been applied to cancers other than breast cancer.

**Comment – 7.8** Study limited to METABRIC, validation in other datasets eg TCGA? Look at where Curtis validated their results.

**Response 7.8 -** Discussed in viva. I agree that it is extremely important to validate our findings in other datasets. In response I have added to future work: “The methods and results in this thesis, researching the role of CNA metrics and allele specific information in prognostic models, has led to a number of interesting outcomes and produced lists of candidate genes or genomic regions from which further exploration, in a research or clinical setting, can be carried out. One of the most important avenues of future work is the validation of these results in another breast cancer dataset, confirming our findings are dataset independent and of potential benefit. The code to run the analyses presented in this thesis is provided on GitHub at: https://github.com/Lydia-King/PhD\_Thesis.”

**Comment 7.9 -** Batch effects in MB gene expression.

**Response 7.9 -** Discussed in viva. In response I have added a sentence to future work to make this clear and emphasize that it will need to be considered in further research: “The gene expression analyses carried out using limma (the survival tree node analysis and CNA state analysis) were performed including only gene expression and CNA state in our models. In addition, no batch correction was carried out. Future work may include application of other gene expression methodologies, e.g. Significance Analysis of Microarrays, expansion to consider other variables of interest, implementation of batch correction techniques, and the possibility of combining gene expression data and our CNA metrics in predictive models for survival outcome.”

**Comment 7.10 -** Alternatives to LIMMA eg samR.

**Response 7.10 -** Discussed in viva and related to comment/response 4.9. In response I have mentioned utilising/exploring alternatives in future work section: “The gene expression analyses carried out using limma (the survival tree node analysis and CNA state analysis) were performed including only gene expression and CNA state in our models. In addition no bacth correction was carried out. Future work may include application of other gene expression methodologies, e.g. Significance Analysis of Microarrays, expansion to consider other variables of interest, implementation of batch correction techniques, and the possibility of combining gene expression data and our CNA metrics in predictive models for survival outcome.”

**Comment 7.11 -** Are events at the genome the full story? Are we still missing Factor X?

**Response 7.11 -** Discussed in viva. Events at genome level are very likely not capturing the full story. Future work could include integrating gene expression data into predictive models to help provide a fuller picture of how these CNA events are impacting function.

**Comment 7.12 -** How can these been applied? What would be anticipated usage? If to be taken forward into clinical use, what would be next steps?

**Response 7.12 -** Next step would be validation in other datasets, mentioned in future work.

**Comment 7.13 -** 'Misexpression of these genes has been documented in literature as facilitating cell proliferation, tumour progression and invasion and as being correlated with survival.' - reference these papers! And discuss in greater detail.

**Response 7.13 -** I have provided references in the conclusion section.

**Comment 7.14 -** ModLim3/5: 'identifying some of the same genes, but also identifying additional genes to be considered for further investigation as candidate biomarkers for breast cancer treatment and outcome.' Talk about these genes in greater depth. Have these been identified in other cancers?

**Response 7.14 -** I have briefly mentioned the possibility of further exploring these genes “identifying additional genes to be considered for further investigation as candidate biomarkers for breast cancer treatment and outcome.”. Lists of these genes are provided in supplementary material and so even though I did not have time to investigate these genes further, they are available for other researchers to look at.

**Comment 7.15** - How would you take this work forward to identify which treatments could be indicated?

**Response 7.15 -** This could be done by analysing a breast cancer dataset that has more up to date and comprehensive treatment information. Using this data, CNA metrics can be produced and compared for patients displaying variation in treatment outcome. This is mentioned in thesis:“In addition, treatment information, including whether a patient received chemotherapy, radiotherapy, hormone therapy, and type of surgery, were not included in our analyses. Inclusion of this information may highlight CNA motifs that confer resistance to certain therapies. Although it should be noted that the METABRIC patients were enrolled between 1977 and 2005 and treatment options and standard of care have changed since e.g. use of Herceptin in HER2+ patients. Therefore, it may be interesting to produce and compare CNA Score and Burden metrics across patients in another more recent breast cancer dataset, for which detailed treatment information is available.”

**Comment 7.16 -** Discuss as to how the results of Chapters 5/6 could be used with respect to treatments/patients?

**Response 7.16 -** Added to thesis **“**These changepoints could potentially provide valuable insights and their impact on survival and treatment response should be investigated in future work.”

**Comment 7.17 -** Triple Neg BC is a disease of unmet need - could your methods be useful with this subtype?

**Response 7.17 -** Yes, triple negative breast cancer is a subtype of breast cancer requiring further research. However, from Chapter 3 it was observed that our CNA metrics were not selected as partitioning variables (useful predictors of survival) for TN breast cancer patients. In contrast, they were useful in stratifying IntClust 3, 4ER+, 7 and 8 patients (in the IntClust trees) or Luminal A and Claudin-Low patients (PAM50 trees) based on survival profiles. There may be potential for our changepoint analysis to provide useful insights in TN breast cancer.

**Comment 7.18 -** How do you know if your models were more effective at predicting survival than Curtis and other prognostic factors? Do you need to compare to existing variables/models?

**Response 7.19 -** We haven't assessed this, mentioned in future work.

**Comment 7.20 -** Look at other CNA/GE datasets, etc those used by Curtis. Identify possible test/validation datasets.

**Response 7.20 -** I recognise that testing and validating the results presented in this thesis is extremely important. See response 7.8.

**Comment 7.21** - Also talk about looking at cross-validation using existing dataset, before taking into unseen dataset. Or staying in METABRIC discovery (including cross-validation), with validation in test set (then further validation in external datasets).

**Response 7.21 -** Cross validation is carried out automatically by survival trees.

**Bibliography:**

**Comment 8.1 -** Check month before year, mostly abbreviated but some numerical months.

**Response 8.1 -** In response I have made sure all months are consistently abbreviated.

**Comment -** Check journal names - all should be abbreviated.

**Response 8.2 -** In responseI have made sure the abbreviations of journals are used in all references.

**Appendices:**

**Comment 9.1 -** Add a short description at the start of each Appendix detailing what it contains.

**Response 9.1 -** In response I have added a short description at the start of each Appendix detailing what it contains.