GENOMIC INSTABILITY OF SOLID TUMORS AND CLINICAL APPLICATIONS

By

Gabrielle W. Choonoo

A THESIS

Presented to the Department of Medical Informatics and Clinical Epidemiology
and the Oregon Health & Science University
School of Medicine
in partial fulfillment of
the requirements for the degree of

Master of Science

December 2018

School of Medicine

Oregon Health & Science University

CERTIFICATE OF APPROVAL

This is to certify that the Master's thesis of

Gabrielle W. Choonoo

has been approved

Dr. Michael Mooney
Mentor/Advisor

Dr. Shannon McWeeney
Committee Member

Dr. Ted Laderas
Committee Member

TABLE OF CONTENTS

List of Tables

List of Figures

Acknowledgments

I would like to thank my thesis advisor, Dr. Michael Mooney, for his mentorship, guidance, and support throughout my thesis project. I would also like to thank my thesis advisory committee members, Dr. Ted Laderas, and Dr. Shannon McWeeney, for providing their leadership in the field, and continual encouragement. I am also grateful for Oregon Health and Science University, and the Department of Medical Informatics and Clinical Epidemiology in the School of Medicine, including the students, faculty and staff. Finally, I am grateful to my parents and extended family for instilling upon me a love for education, and also my sibling and friends for always believing in me.

Abstract

Chromosomal abnormalities caused by genomic instability are a consistent feature of cancer cells. However, the mechanisms that lead to genomic instability and its role in cancer progression, and ultimately clinical outcome, are not well understood. Advances in whole genome sequencing and other high-throughput molecular techniques have vastly improved our ability to characterize cancers. Researchers have begun to measure genomic/chromosomal instability (GI/CIN) by using multiple data types, including copy number, gene expression, and somatic mutations. These measures of instability have been associated with various clinical outcomes, but results have been inconsistent. Several methods for creating summary measures of CIN have been published, but there is no consensus about the best method. Additionally, it is unclear whether any one method is appropriate for all cancer types. Current methods also do not take advantage of information about specific cytogenetic abnormalities that are recurrent in specific cancer types. Given the potential clinical value of CIN as a predictor of cancer progression or treatment response, there is a critical need to better understand how best to summarize measures of CIN.

To address this gap in the research, I annotated the CIN, GI, and GI recurrent scores for four tumor types, breast cancer (BRCA), ovarian cancer (OV), kidney cancer (KIRC), and colon cancer (COAD) from The Cancer Genome Atlas (TCGA) database and performed univariate and multivariate survival analysis of the scores. The findings showed that chromosomal instability is, in fact, specific to tumor type and that the clinical outcomes were not consistent. No single score predicted all cancer types equally well. The overall, univariate survival analysis revealed that CIN25 scores were predictive of breast and kidney cancer, where a lower score was associated with better outcomes. Similarly, the GI score was predictive of breast, kidney and ovarian cancer. However, the outcomes for the ovarian cancer were the opposite, where a lower score was associated with worse outcomes. The scores were not significantly predictive of the colon cancer dataset. These findings highlight the continued importance of cancer-specific scores and appropriate evaluation to assess correct measures for each disease type. The modified GI scores that contained recurrent abnormality regions supported the findings from the GI score and provide a beginning framework for a cancer-type specific score that incorporates both structural abnormalities as well as genomic aberrations.

Chapter 1: Introduction

1.1 What is Genomic Instability?

By definition, genomic instability refers to a genetically defective state, which stems from the manifestation of mutations and chromosomal rearrangements, or abnormality [1]. Genomic variation exists on a scale of the number of nucleotides. At the smaller end of the scale, we have mutations, which involve one nucleotide change. Whereas, at the larger end of the scale, we have chromosomal instability and cytogenetics, which can include 10^8 number of nucleotides. Why should it be important to study genomic instability? Genomic instability is a hallmark of all cancer types, meaning that much of what we know about the mechanism of cancer, treatment, and survival, is a consequence of genomic instability.

1.2 Clinical Decision Making

Because genomic instability is a hallmark of cancer, it has been used for clinical decision making, particularly on the larger end of the scale. The clinical utility on this larger end of the scale includes discussions about gene fusions and chromosomal instability, also referred to as clinical cytogenetics. Clinical cytogenetics can include structural changes that are routinely used for clinical decision making, particularly for hematologic malignancies. However, for solid tumors, the clinical utility is still uncertain. Researchers have used information from different scales to summarize instability, however the clinical outcomes have often been inconsistent.

Depending on the cancer type, different types of genetic variation have been useful for clinical decision making. The most known cytogenetic example in cancer is the Philadelphia chromosome, which is a reciprocal translocation of chromosome 9 to chromosome 22, found in chronic myeloid leukemia patients. The result is a fusion gene known as BCR-ABL1, which causes uncontrollable cell growth. This fusion gene discovery has been used for clinical diagnosis and prediction of patient outcomes [2]. Whereas for breast and ovarian cancer, subtyping based on germline mutations, such as breast cancer 1 (BRCA1) and breast cancer 2 (BRCA2), or the overexpression of the human epidermal growth factor receptor 2 (HER2), can affect clinical decision making.

1.3 Chromosomal Abnormality Structures

Addition, deletion, substitution, and translocation – these are the most known structural chromosomal abnormalities that can be evaluated for clinical decision making. These structural rearrangements can involve whole chromosomes, or pieces of chromosomes. Whole chromosomes can get deleted or duplicated, or this can occur within sections of a particular chromosome. However, substitution and translocation can get more complicated by involving more than one chromosome, where pieces of different chromosomes get interchanged [3].

Historically, these chromosomal structural changes were studied by isolating, staining, and viewing chromosomes under a microscope. This microscopic picture of chromosomes is called a karyotype. Cytogenetic analyses include Fluorescence in situ hybridization (FISH), which maps specific DNA sequences on chromosomes and has increased the efficiency and accuracy of cytogenetic analysis. There are several variations of this method that provide higher resolution. More recent methods include high throughput genomic methods such as array based comparative genomic hybridization and next generation sequencing [4-5].

1.4 Hematologic Malignancies versus Solid Tumors

Chromosome abnormalities have been detected in hematologic malignancies. Typically, about half of the patients with acute leukemia, including both myeloid and lymphoid, have normal karyotype in their leukemic cells. One of the most known examples is the Philadelphia chromosome (Ph), which is used as a diagnostic and prognostic indicator for patients with chronic myeloid leukemia. Chromosome abnormality studies have become part of the diagnostic, therapeutic, and prognostic measures in hematological malignancies [6]. Cytogenetics have also been combined with genomic landscape studies in which the overall cytogenetic risk is assessed. For example, miRNA-sequencing data for acute myeloid leukemia was subtyped and associated with different cytogenetic risk categories [7].

There are some technical challenges in achieving karyotypes for solid tumors, which has hindered the research of chromosomal abnormality in solid tumors compared to hematologic malignancies. Damage of cancer cells before culturing is what causes some of these technical challenges. Also, tumor progression causes more complex chromosomal changes compared to hematologic malignancies. These complexities make it difficult to identify primary chromosome changes for each solid tumor type [6].

Despite these challenges, the FISH technique has led to advances in chromosomal abnormality research in solid tumors because it can be utilized on fresh tumor tissue, exfoliative cells, and embedded specimen [6]. For example, overexpression of HER2 is known to be associated with poor survival, yet sensitive to targeted therapy, trastuzumab [8]. However, researchers found that the FISH assay using HER2 and chromosome 17 sequences was a more accurate and reliable way to predict response to therapy, rather than evaluating HER2 protein expression alone [8]. In contrast, Das et. al. found that a combination of FISH techniques was more accurate in predicting prognosis of myeloid malignancies [5]. In summary, these findings illustrate that clinical applications can potentially be affected by the particular methods, or combination of methods, that are used to discover chromosomal abnormalities.

In addition to the diverse methods used for discovering chromosomal abnormalities, researchers have found that the role of cytogenetics depends on the tumor type [2]. For example, cytogenetics was used to help discover screening and response to targeted therapy in lung cancer, while diagnosis and prognosis was concluded from cytogenetic abnormalities in cutaneous and uveal melanoma. In summary, the clinical utility may be specific to each tumor type. Instability measures have the potential to be assess clinical outcomes across different tumor types.

1.5 Chromosomal Instability Scores

Because traditional methods for discovering chromosomal abnormality are not routinely used for clinical decision making in solid tumors, the data is scarce. Therefore, scores were developed to attempt to measure abnormality, and in turn, measure different kinds of instability. Chromosome instability (CIN) and genomic instability (GI) scores attempt to provide a summary measure of overall instability using gene expression, copy number, and somatic mutation. Although many tumors show chromosomal abnormalities, CIN is characterized by an increased rate of these errors [9]. Specifically, CIN uses gene expression to represent instability at the chromosome level, whereas GI uses both mutations and copy number variations to encompass different scales of instability. CIN25/70, GI, and CINdex – these are the three different methods I planned to implement from previous studies that capture genomic or chromosome instability.

The CIN25/70 method is based on a 70-gene, and 25-gene, expression signature that incorporated different tumor types to predict prognosis [10]. Genes were selected based on correlation with functional aneuploidy, which was based off a t-statistic of normalized gene expression for each cytoband region, defined as a particular location on the long or short arm of a chromosome. Each gene was ranked from based on the correlation across different tumors such as breast, ovarian, lung, and prostate. They found that the CIN25/70 score performed well for most of the datasets, where lower scores were associated with a higher rate of survival. However, they found that the scores did not have significant results in predicting survival for ovarian and some of the lung datasets. In summary, there have been inconsistent clinical findings associated with CIN scores of solid tumors.

The genomic instability (GI) score is calculated using copy number regions and somatic mutations within a cancer genome [11]. The equation is K*n1 + n2 for each patient, where K was set to 0.5 because this significantly discriminated between long and short median overall survival. N1 is the number of copy number regions and N2 is the number of somatic mutations within a cancer genome. They found that higher GI scores were associated with higher survival rate in ovarian cancer, regardless of BRCA1/2 mutations. Multivariate analyses revealed that the score was independently prognostic for overall survival and progression-free survival. In summary, the CIN and GI scores had opposite outcomes for different tumor types. CIN scores predicted lower scores were associated with better outcomes, whereas GI scores predicted higher scores were associated with better outcomes.

The CINdex score is based on copy number and is available through the R Bioconductor package [12]. The paper discussed the calculation of the CIN score, which was designed to take into account both gains and losses without canceling each other out, and therefore a higher CIN score would be representative of both copy number variations. They calculated this score for gains, losses, and the combination of both. They found that thresholds of 2.25 and 1.75 worked the best for displaying heatmap resolution of each sample's score across each chromosome comparing relapse versus non-relapse patients. A higher loss CIN score was associated with better prognosis because almost the entire chromosome 20 was lost in the relapse-free group. Although I planned on using this instability measure, the software did not have the option of using the calculation described in the paper. For this reason, I chose to focus on method extension for the remaining the GI scores.

The method extension for the GI score involves incorporating the recurrent abnormalities into the score. This modification of the GI score is an extension of the method used from the previous study because it combines both structural chromosomal changes and genomic aberrations. These recurrent regions are defined as cytoband regions that contain structural chromosome abnormalities that frequently occur in specific tumor types. The rationale for this is that there is evidence that genomic instability in specific regions are more predictive of outcomes. Also, this method is a first attempt at modifying the GI score to create a cancer-type specific score that could potentially predict outcomes more accurately than clinical measures.

Overall, these chromosome instability scores measure different types of instability and are associated with different clinical findings. CIN25/70 is summarized at the overall transcriptional level, and the GI score is based on sample specific aberrations along with large structural variation. The CIN25/70 score predicted higher CIN score would have worse prognosis, but did not perform well for all datasets. In contrast, the GI score found that higher CIN score was associated with better outcomes.

1.6 Specific Aims

Because clinical outcomes have been inconsistent regarding chromosomal instability in solid tumors, my research motivation was to understand the best way to summarize measures of

CIN. Several methods for creating summary measures of CIN have been published, however there is no consensus about the best method to use. The current methods do not take into account specific cytogenetic abnormalities that are recurrent in specific cancer types. There is a potential clinical value of CIN as a predictor of cancer progression or treatment response. My hypothesis is that using combinations of current chromosome instability measures, which utilize different data types, will provide greater ability to predict outcomes across a wide variety of cancers. The motivation is to provide evidence for how best to use measures of instability to predict clinical outcomes and also to identify which measures are the most predictive of clinical features in a variety of solid tumors.

For Aim 1, I developed comprehensive annotations related to genomic instability for TCGA samples. The scores integrate multiple data types and instability scores that apply to four cancer types available through TCGA including breast cancer (BRCA), ovarian cancer (OV), colon cancer (COAD), and kidney cancer (KIRC). Specifically, the CIN25/70 uses the 70-gene, and 25-gene, expression signature, the GI score uses the copy number and somatic mutation, and the modification of the GI score is based off recurrent abnormalities in solid tumors. The outcome of aim 1 provides case-by-case annotations as well as a way to calculate a cancer-type specific score, which could potentially better predict since it incorporates recurrent abnormalities relative to each tumor type. There were a total of 12 tumor type and score comparisons. The breast and ovarian tumor types were utilized in the previous CIN25 study and the ovarian tumor type was assessed in the GI study, however these datasets did not use the most recent TCGA data. The remaining 9 tumor type and score comparisons were not previously studied and so these provide new findings to the field.

For Aim 2, I investigated the predictive ability of the genomic instability annotations across a variety of cancer types. I performed univariate and multivariate survival analyses for each of the scores to determine clinical outcomes. The survival outcomes I analyzed were overall survival, progression free survival, and tumor stage. The results of these analyses will identify clinically relevant outcomes such as differences in clinical outcomes for diverse cancer types with regard to instability scores, whether scores predict outcomes better than common clinical measures, and whether scores predict outcomes in specific patient cohorts such as early or late stage tumors. The outcome of aim 2 will also identify methodological outcomes such as how combining current chromosome instability measures, by using different data types, affects the prediction of clinical outcomes in a variety of cancers, and also how to incorporate recurrent chromosomal abnormalities into these measures.

The results of this thesis will potentially add to the clinical value of genomic instability in cancer research. This method will provide case-by-case analyses on the genomic instability data of solid tumors and understand the predictive power of instability scores. This will provide a greater understanding of the genetics driving solid tumors and ways to improve care for patients.

1.7 Mapping of the Document

In Chapter 2, I describe the methods used to process TCGA clinical and genomic data, annotate the chromosomal instability scores using genomic data, replicate previous studies, and exploratory data analysis of the scores stratified by clinical features. Then, in Chapter 3, I discuss the results of the survival analyses of the scores across four tumor types. In Chapter 4, I discuss the overall findings, recommendations, and significance to the field. Lastly, in Chapter 5, I provide an overall summary of the work, future work, and limitations of the study.

Chapter 2: Material and Methods

After an IRB determination inquiry, this study was exempt from IRB review and approval. In this chapter, I discuss the methods I used to gather the clinical and genomic TCGA for breast cancer (BRCA), ovarian cancer (OV), colon cancer (COAD), and kidney cancer (KIRC). Then I describe the methods I used to annotate the CIN and GI scores, as well as the modified GI score. Finally, I discuss the process of replicating previous studies and exploratory data analysis of the scores stratified by clinical features, as well as the survival analysis methods.

2.1 Processing Clinical data

Over 60 tumor types exist in The Cancer Genome Atlas (TCGA) data, available through the National Cancer Institute's Genomic Data Commons (GDC) portal. The aim of the GDC database is to provide standardized cancer genomic data in support of precision medicine within the cancer research community [13]. For the clinical data, I searched each tumor type separately, and added all the XML files available for each patient within the study to my cart. Then I downloaded the manifest file for my cart and used the gdc-client in Linux to download the files. Once the XML files were downloaded, I used a Python script that I coded to parse the XML files into text files. Finally, I used R to read in the parsed XML files to format in a table, where the columns were clinical variables, and rows were each patient.

By definition, tumor progression occurs if a patient is diagnosed with a secondary tumor, which can occur if a primary tumor has metastasized or recurred after surgery. Because much of the clinical data has missing data or discrepancies among progression indicators, I used several indicators to classify progression status. Patients were annotated as progressors if they had at least one occurrence of days to new tumor, new tumor event, new tumor site, progressive status, or metastatic site. I also took note of the patients that had sequenced recurrent or metastatic tissue samples and annotated these as progressors as well. Patients were annotated as nonprogressors if they did not have any occurrences of days to new tumor, no new tumor event, no new tumor site, remission status, no metastatic site, and were not annotated as a progressor. Furthermore, if a progressor had a new tumor event annotated as "no", but also had days to new tumor, then I annotated them to have a new tumor event.

Once the progression status was annotated, there were a few more data cleaning procedures for number of days and tumor stage. If a patient had more than one days to new tumor, days to last follow-up, or days to death, then I used the maximum number of days. Tumor stages I and II were annotated as early stage tumors, and stages III to IV were annotated as late stage tumors. Sublevels within numerical tumor stages, such as A, B, and C, were included in the same numerical category. For example, stage IA, IIB, and IIC were all categorized as early stage tumors.

Data coverage in TCGA is not always uniform, meaning that it is possible for patients to have genomic data, but not clinical annotation, and vice versa. The number of patients with clinical data also varies across tumor types. There were a total of 995 BRCA patients with clinical data, 459 COAD patients, 587 OV patients, and 537 KIRC patients. These sample sizes decrease for the survival analysis since not all patients necessarily have genomic data used to calculate chromosome instability scores. In the next section, I will discuss the genomic data cleaning process and sample sizes.

2.2 Processing Genomic Data

Section 2.2 covers the methods I took to download and parse the genomic data used to calculate the CIN and GI scores from previous studies. These were annotated for each patient across the four tumor types. For all data types, I used the most recent genome reference consortium human build, 38 (GRCh 38). For this study, I only used masked data for copy number and mutation, meaning that germline mutations were not included. Lastly, I only used primary tissue samples.

2.2.a Downloading and parsing TCGA genomic data

Gene expression, copy number, and somatic mutation – these are the genomic data types I downloaded from TCGA. Similar to the clinical data, I searched each tumor type separately in the GDC commons database, add the gene expression files to my cart, and downloaded the manifest file. Using Linux, I used the gdc-client function to download the manifest file. The expression files for each patient get saved to a separate folder. Then I use an R script I coded to download all the files, match them to their patient ID, and format the data. Gene expression data was formatted for each tumor type in which rows were genes, columns were patient IDs, and each cell contained the gene expression value. TCGA has three different gene expression formats known as Hit Seq, which are the raw counts, FPKM which are the normalized counts that account for gene length and the number of reads mapped to all protein-coding genes, and FPKM upper-quantile, which is a modified version of FPKM where the 75th percentile is used as the denominator in place of the total number of protein-coding reads. For this study, I used the FPKM normalized counts.

The TCGA copy number pipeline uses Affymetrix SNP 6.0 array data and the DNAcopy Bioconductor package to perform circular binary segmentation analysis [14]. For this study, I used the masked copy number segments only, meaning that the data went through a filtering process to remove known probes containing germline mutations. I used a similar method to process the copy number that I used for the gene expression data. The cleaned format, however, was different because the rows were patient IDs and each column represented the copy number segment such as genomic start position, stop position, chromosome, and segment mean. Lastly, I annotated the region size, which is the end position minus the start position.

TCGA has four methods for calling somatic mutations known as Mutect2, Muse, VarScan2, and SomaticSniper and they are saved in mutation annotation format (MAF). For this study, I used mutect2 because it has a low false positive rate [15]. I downloaded the associated MAF file for each tumor type, filtered primary tissue samples, and high impact mutations. Mutation impact are annotated from VEP, where high impact denotes the variant is assumed to have disruptive impact in the protein, probably causing protein truncation, loss of function or triggering nonsense mediated decay [16].

2.2.b Downloading and parsing Mitelman data

The Mitelman Database contains recurrent chromosome aberrations across a variety of tumor types and is manually curated by literature from Felix Mitelman, Bertil Johansson, and Fredrik Mertens. Recurrent chromosome aberrations are defined as structural or numerical abnormalities that are present in at least two patients with the same morphology or topography [17]. These structural abnormalities include deletion, substitution, translocation, etc. on specific cytobands.

From the online database, I downloaded all balanced and unbalanced structural recurrent aberrations for breast, ovarian, large intestine, and kidney topographies (Time stamped August 13th, 2018). For each dataset, I filtered for the morphologies that were adenocarcinoma. The patients, or subjects, included in the database are referred to as “cases”. I downloaded the individual cases for each tumor type to get the total number of cases that were evaluated. To annotate the percentage of cases that each abnormality existed in, I added a column in the recurrent dataset that was the number of cases divided the total number of cases. In choosing the recurrent abnormalities to be used for modifying the GI score, we decided to include all recurrent abnormalities, as well as the abnormalities that were present in more than 20% of the cases, and more than 10 cases. The reason for the two thresholds is the difference in total number of cases for each tumor type. The breast cancer dataset had 138 cases, the ovarian dataset had 36 cases, the colon dataset had 60 cases, and the kidney cancer dataset had 131 cases.

2.3 Annotating CIN and GI Scores

Once the genomic data was downloaded and cleaned, I was able to use the data to annotate the CIN and GI scores. First, I downloaded the 70-gene signature from paper by Carter et. al. [10] and mapped these gene symbols to Ensembl gene names using the Biomart R package [18-19]. Then I filtered the gene expression data to include only these 70 genes. For each patient, I calculated the sum expression across the 70 genes for the CIN70 score, and annotated patients with scores greater than or equal to the median as high, and less than the median as low. In addition, I calculated the CIN25 score the same way, except only using the top 25 ranked genes.

For the GI score, I followed the filtering process used in the paper by Zhang et. al., which included copy number segments that were in regions greater than 3Mb, and had segment values greater than 0.05, and less than -0.05 [11]. To remove noise, I filtered primary samples that had normal copy number segments that existed in the same region, met the same criteria, and were in the same direction as the primary samples of either positive or negative. Then I calculated the number of copy number regions and the number mutations to calculate the GI score for each patient: 0.5*copy number regions + mutations. The scores were again differentiated by the median across patients, where scores greater than or equal to the median were categorized as high scores, and less than the median were categorized as low scores.

For the modified GI score based on recurrent abnormalities, I filtered the genomic regions of the copy number and mutation to those that existed on the cytobands with recurrent abnormalities for each tumor type. I used the Biovizbase R package [20] to retrieve the reference human genome 38 cytobands. Then I merged this with the copy number data, based on the chromosome, and annotated the cytobands for each copy number regions based on if the region was completely contained, or overlapping, with the reference cytobands. Once I mapped the copy number regions to cytobands, I filtered the regions that were contained on recurrent abnormalities specific to each tumor type. I did the same thing for the normal samples, and filtered the primary samples, similar to the original GI score. Similarly, for the mutations, I only included the mutations contained within or overlapping with recurrent cytobands. I then used the same GI score equation and categorized the scores by the median across patients.

2.4 Replicating previous studies

To replicate the previous studies, I downloaded the datasets that were used in the papers and computed survival analysis to assure that my method for annotating the scores was correct. Specifically, I used the breast cancer dataset from the Carter et. al. paper that was originally adapted from the study by Sotiriou et. al. [21]. In addition, I used the archived TCGA ovarian data set that was used in the study by Zhang et. al [11]. There were some limitations with replication due to not being able to get the exact data that was used, however I was able to use similar data based on the method descriptions in the papers. For replication, I used the mean instead of the median since that is how they classified them in the paper by Carter el. al. For the TCGA datasets, I used the median to be consistent with the GI scores.

2.4.a CIN25/70 scores

GSE2990, a dataset available in the Gene Expression Omnibus (GEO) data repository, is the breast cancer expression array I used to replicate the findings in the Carter el. al. paper, which found that a higher CIN25/70 score was associated with a lower rate of survival based on a Kaplan Meir survival curve [10, Figure 1]. To replicate these findings, I downloaded the array, and read in the CEL files using the Bioconductor package, oligo [22]. Then I normalized the data using the robust multichip average algorithm and filtered multiple probes for one gene based on the largest test statistic using the genefilter Bioconductor package [23]. Lastly, I annotated the CIN25 and CIN70 score for patients and computed the survival curve. I achieved the same results from the paper, where a higher score was significantly associated with a lower rate of survival in breast cancer patients, with the same sample size of 179 patients (Figure 2).

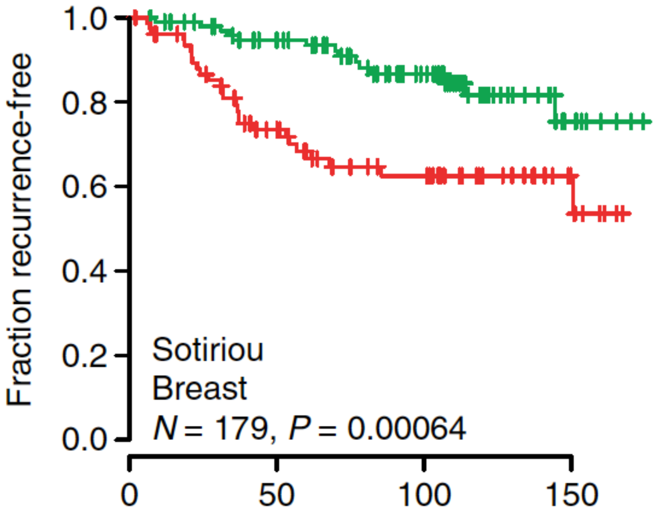


Figure 1. Actual results of the breast cancer recurrence-free survival results of CIN25 scores ($p < 0.05$) where red

indicates high score, or scores that are above average, and green indicates low score, or scores that are below average. Significant results indicate that higher CIN25 is associated with poor prognosis as seen in the Carter et. al. paper, Figure 2, row 3, column 3 [10].

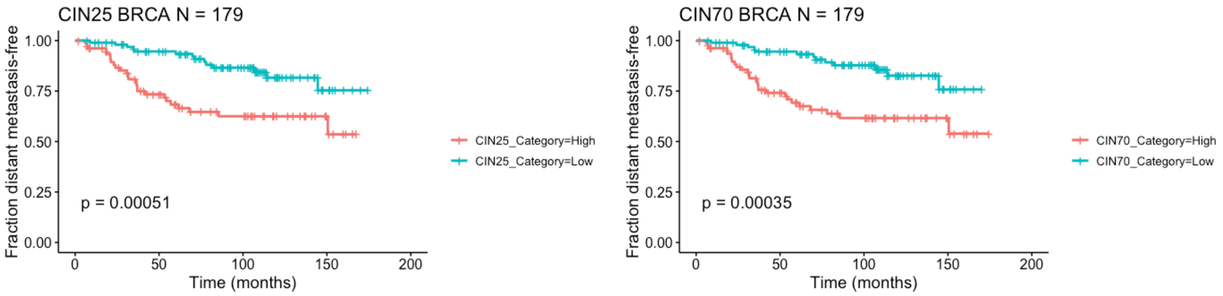


Figure 2. Replication results of the breast cancer distant metastasis-free survival results of CIN25/70 scores ($p < 0.05$). Significant results indicate that higher CIN25/70 is associated with poor prognosis.

2.4.b GI scores

For the GI score, I downloaded the archived TCGA ovarian data including the clinical data, copy number and mutation, however there were some differences in sample sizes between the archived version and the version used in the paper. I used the valid, whole exome sequencing, build 37 mutations, following the methods in the Zhang et.al. paper [11]. However, the number of mutations in the paper was 14,970, whereas the archived data had slightly less: 14,956 mutations. I used the same method as the paper by filtering copy number regions greater than 3Mb, and segment values less than -0.05 or greater than 0.05. The number of patients with mutations in the archived dataset was 321, and of these, 312 also had both copy number and clinical data. In contrast, they had a sample size of 325 that had all three data types. There was also a difference in sample size for the progression-free survival due to the fact that more follow-up data had been archived since the study was published. This gave 372 patients with progression-free survival data, instead of 325 as seen in the paper.

Although there were some differences in sample size, the survival outcomes were the same. The overall survival was calculated using the number of years to death, and the vital status of the patient. Progression-free survival was calculated using the number of years to follow-up, and if the patient had a new tumor event. Figure 3 shows the results from the paper, where a higher GI score was significantly associated with better overall survival and better progression-free survival. The multivariate analysis shows that the scores independently predicted outcomes, and were not confounded with age, tumor grade, debulk, and stage. Figures 4a and 4b shows the univariate replication results, where I also found that, although the samples sizes and p-values were slightly different, a higher GI score was significantly associated with better overall survival and better progression-free survival. Figures 4c and 4d show the scores were not significant in the multivariate model, however they follow the same trend that lower score is associated with worse prognosis.

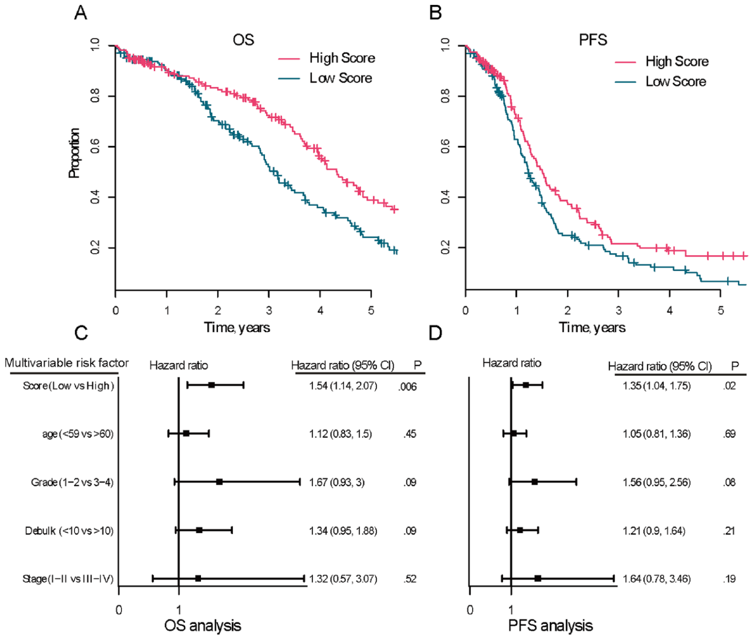


Figure 3. Actual results of the archived TCGA ovarian data overall survival and progression-free survival for the GI scores ($p < 0.05$). Significant results indicate that higher GI score is associated with good prognosis as seen in the Zhang et. al. paper, Figure 4A and 4B [11]. Their analysis had 325 patients, with log rank p=0.004, and p=0.009,

respectively.

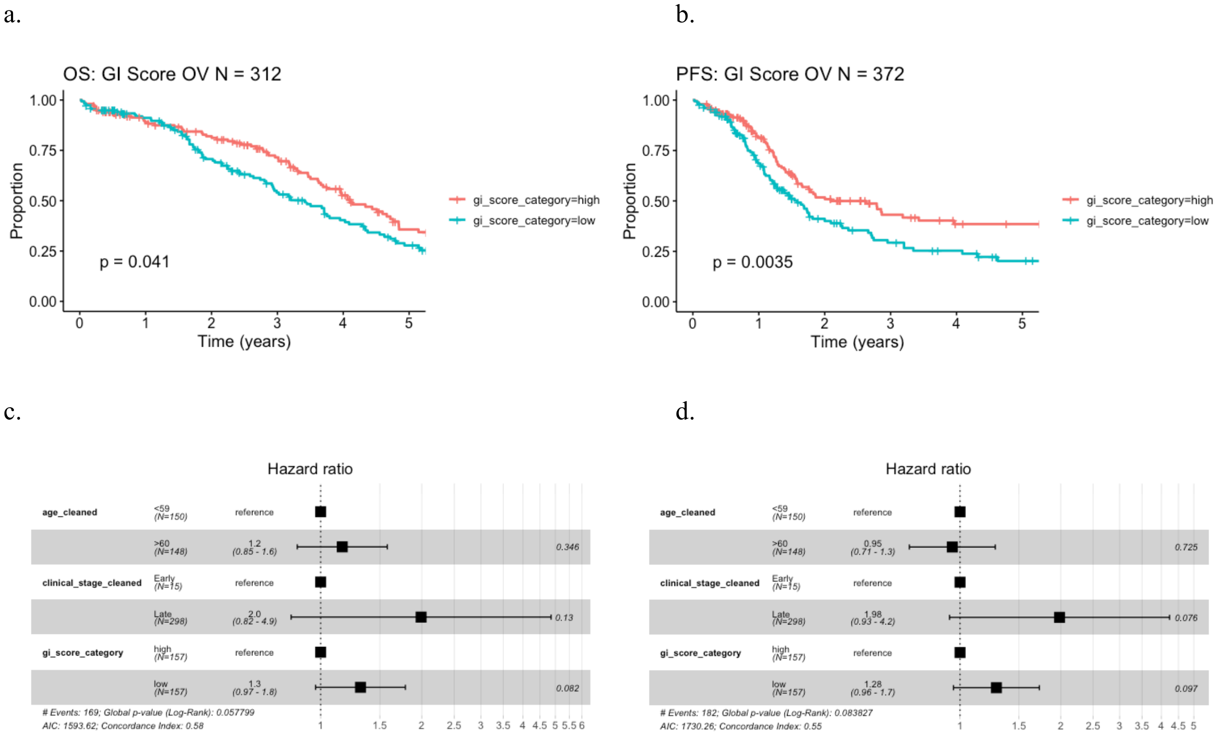


Figure 4. a-b) Replication results of the archived TCGA ovarian data. Univariate analysis reveals overall survival and progression-free survival for the GI scores are significant, where higher GI score is associated with good prognosis as seen in the Zhang et. al. paper, Figure 4A and 4B [11]. This analysis had 312 patients for the overall survival, and 372 patients for the progression-free survival, with p-values 0.041, and 0.0035, respectively. c-d) Replication results of the multivariate analysis for overall survival (c) and progression-free survival (d). Results do not have significant p-values, but the trend is the same as the paper where a lower GI score is associated with worse prognosis.

2.5 Exploratory Data analysis

After I replicated the results from the papers, I used the methods described in section 2.3 to annotate the current TCGA data for four tumor types, breast, ovarian, colon, and kidney, and performed exploratory data analysis (EDA) of the scores. Note that I removed two outliers that had more than 1000 mutations from the COAD and KIRC dataset. I assessed correlations between the different instability scores, score distributions across tumor types, and scores stratified by clinical features. The main EDA plots are in this section and additional plots can be found in Appendix 1. In summary, I found that the scores are significantly correlated, but that the scores are providing different information. I also found that both standardized and raw scores had significantly different distributions across tumor types. Lastly, for the clinical data, there were some significant differences across tumor type scores stratified by age and tumor stage.

As expected, the CIN25 and CIN70 were highly correlated with a p-value near 0, using Pearson's correlation coefficient [24], and so the CIN70 scores are not utilized in the analysis. The CIN25 and GI scores were also significantly correlated to the GI scores; however, the linear relationships were not exactly the same, indicating that the CIN and GI scores were providing different information (Appendix 1.1). The strength of the correlation between the CIN scores and GI scores were modest, ranging from 0.19 to 0.5. BRCA had the highest correlation between the GI score and the CIN25 score at 0.5, with a p-value near 0. OV, KIRC, and COAD had correlations around 0.24, 0.33, and 0.19, respectively (Table 1).

Table 1. Correlation coefficients and p-values between each score across tumor types. All correlations are significant at the p<0.05 level.

Tumor Type	CIN25-GI Score Correlation/p-value	CIN25-GI Score Recurrent Correlation/p-value	GI Score-GI Score Recurrent Correlation/p-value
BRCA	0.50/9.2e-39	0.51/7.9e-42	0.99/0
OV	0.28/0.0002	0.24/0.002	0.98/5.4e-116
KIRC	0.32/7.1e-08	0.33/6.1e-08	0.99/3.3e-231
COAD	0.19/0.0006	0.19/0.006	0.99/6.3e-302

Figure 6 shows the boxplot distributions of the standardized scores across tumor types. Scores were standardized by computing a z-score of each score across all tumor types. The raw distributions can be found in Appendix 1.2. An analysis of variance (ANOVA) revealed that both standardized and raw scores had significantly different distributions across tumor types, but not across score types (Table A1.3-4). From the distributions, it can be observed that the KIRC dataset scores is significantly lower than the other tumor types for each score, and the BRCA and OV datasets are fairly similar (Figure 6).

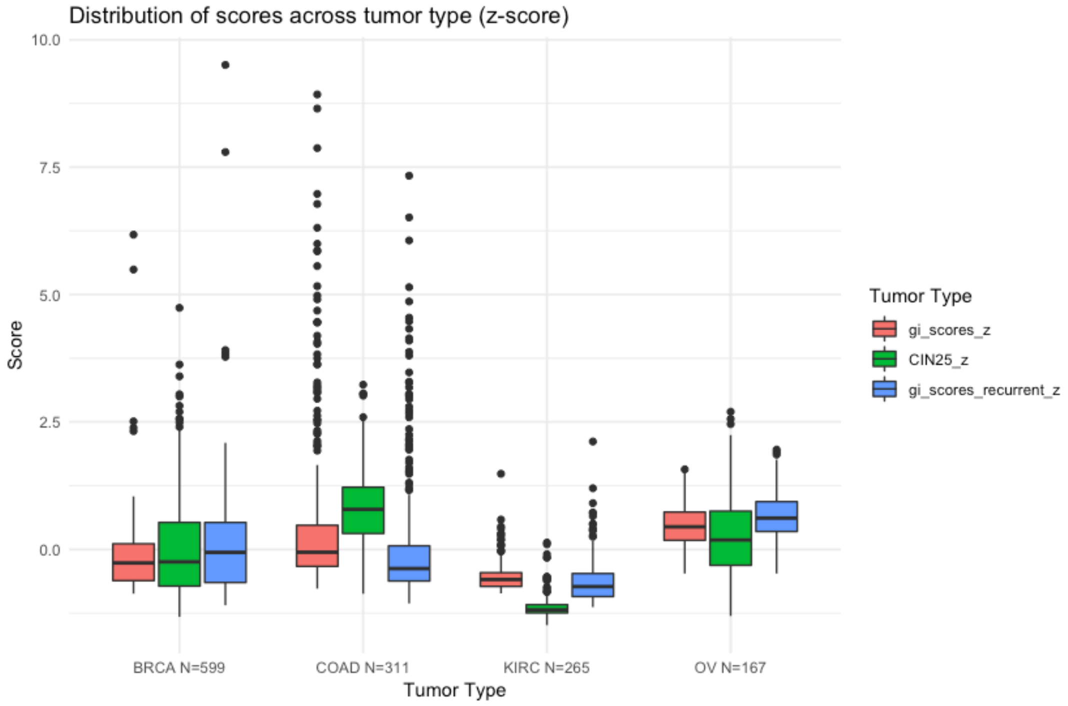


Figure 5. Distributions of standardized scores across tumor types. The x-axis shows the four different tumor types. Within each tumor type, the boxplots show the distribution of the three different scores, GI scores, CIN25, and GI scores recurrent, where pink is the GI score, green is the CIN25 score, and blue is the GI recurrent scores.

For the clinical data, there were significant differences in the overall age at diagnosis (dx), gender, vital status, new tumor event, tumor stage, and progression across each tumor type

(Appendix 1.5-1.10). Specifically, the COAD age at dx was significantly higher than the BRCA, OV, and KIRC datasets. Further, the KIRC dataset was more uniformly distributed that the BRCA dataset. Not surprisingly, the BRCA dataset had significantly more females than males. For the remaining overall comparisons, the OV dataset had significantly more deceased patients, patients that had a new tumor event, late stage tumors, and progressors (Table 2).

Table 2. Overall summary of clinical variables for each tumor type.

Tumor type	Age at dx	Gender	Vital status	New tumor event	Tumor stage	Progression

BRCA	26-90 (58)	99% Female;
		1% Male
			86% Alive;
			14% Dead
				70% No;
				13% Yes;
				17% NA
					74% Early;
					25% Late
						84% NonProgressors;
						16% Progressors

OV	26-89 (60)	100% Female	40% Alive;
			60% Dead
				2% No;
				30% Yes;
				68% NA
					8% Early;
					91% Late;
					1% NA
						26% NonProgressors;
						74% Progressors

COAD	31-90 (67)	47% Female;
		53% Male
			78% Alive;
			22% Dead
				43% No;
				42% Yes;
				15% NA
					56% Early;
					42% Late;
					2% NA
						44% NonProgressors;
						56% Progressors

KIRC	26-90 (61)	36% Female;
		64% Male
			67% Alive;
			33% Dead
				21% No;
				6% Yes;
				73% NA
					61% Early;
					39% Late
						67% NonProgressors;
						33% Progressors

In addition to overall comparisons, I stratified the clinical variables of age at diagnosis, gender, and tumor stage by score to compare distributions (Appendix 1.11-1.19). In comparing the clinical features stratified by score, mean age at dx was significantly higher for KIRC patients with high GI scores compared to KIRC patients with low GI scores. In contrast, mean age at dx was significantly lower for BRCA patients with high CIN25 scores compared to BRCA patients with high CIN25 scores. There were no significant differences in gender stratified by scores. However, for tumor stage, the KIRC dataset had significantly more early stage tumors with low GI and CIN25 scores (Table 3).

Table 3. Overall summary of clinical variables for each tumor type by score.

Tumor type	Age at dx	Gender	Tumor stage
BRCA	58 High; 58 Low	99% Female High; 1% Male High; 99% Female Low; 1% Male Low	73% Early High; 27% Late High; 79% Early Low; 21% Late Low
OV	60 High; 60 Low	100% Female	8% Early High; 92% Late High; 5% Early Low; 95% Late Low
COAD	67 High; 65 Low	52% Female High; 48% Male High; 46% Female Low; 54% Male Low	55% Early High; 45% Late High; 55% Early Low; 45% Late Low
KIRC	63 High; 57 Low	36% Female High; 63% Male High; 36% Female Low; 63% Male Low	56% Early High; 44% Late High; 70% Early Low; 30% Late Low

2.6 Survival Analysis Methods

For the survival analyses, I performed both univariate and multivariate analysis using the survival R package [25] to assess if individual scores had different effects on survival versus combined scores with other cofactors such as age, gender, and tumor stage. The univariate analysis involved the Kaplan-Meier model and the multivariate analysis involved the Cox proportional hazards regression model. Both of these models fit survival curves based on a time and event factor as the outcome stratified by covariates. For the univariate analysis, I computed overall survival using years to death for deceased patients and years to follow-up for living patients. Furthermore, I computed progression-free survival using years to new tumor event for patients that had a new tumor and years to follow-up for patients that did not have a new tumor. I added covariates such as age, gender, tumor stage, and included both scores to the multivariate model. CIN25 was used in the analysis, instead of CIN70, since the scores were highly correlated and this was the focus in the Carter el. al. paper. The previous papers used a 5-year survival, however I computed survival out to maximum that the patient was followed.

Chapter 3: Results

3.1 Survival Analysis

For the survival analysis, I assessed both overall survival and progression-free survival. For the univariate overall survival analysis, I assessed the outcomes using days to death for deceased patients, and days to last follow-up for living patients, and used scores as predictors. The score predictors were categorical variables defined as scores greater than or equal to the median as high scores, and less than the median as low scores. For the univariate progression-free survival, I assessed the outcomes using days to new tumor for patients that had a new tumor event, and days to last follow-up for patients that did not have a new tumor, and used scores as predictors. The multivariate analyses used age, gender, and tumor stage as covariates to assess any confounding factors in the model.

3.1.a Univariate analysis

For the overall survival using the CIN25 score, the BRCA and KIRC datasets were significant and consistent with findings from the Carter et. al. paper that a lower score is associated with better outcomes (Figure 6). In contrast, the OV and COAD datasets were not significant. Neither the colon cancer dataset or the kidney cancer dataset was implemented in the Carter et. al. paper [10]. However, the results reflect the findings from the paper where BRCA had significant results, and the OV dataset did not. In summary, this indicates that scores have different predictive value depending on tumor type. The CIN25 score can be used to predict the outcomes for the BRCA and KIRC dataset, however it does not predict the outcomes for the

COAD and OV dataset. Also, note that the trend drops off after about 10 years for the BRCA dataset, indicating that the score is not predictive after 10 years.

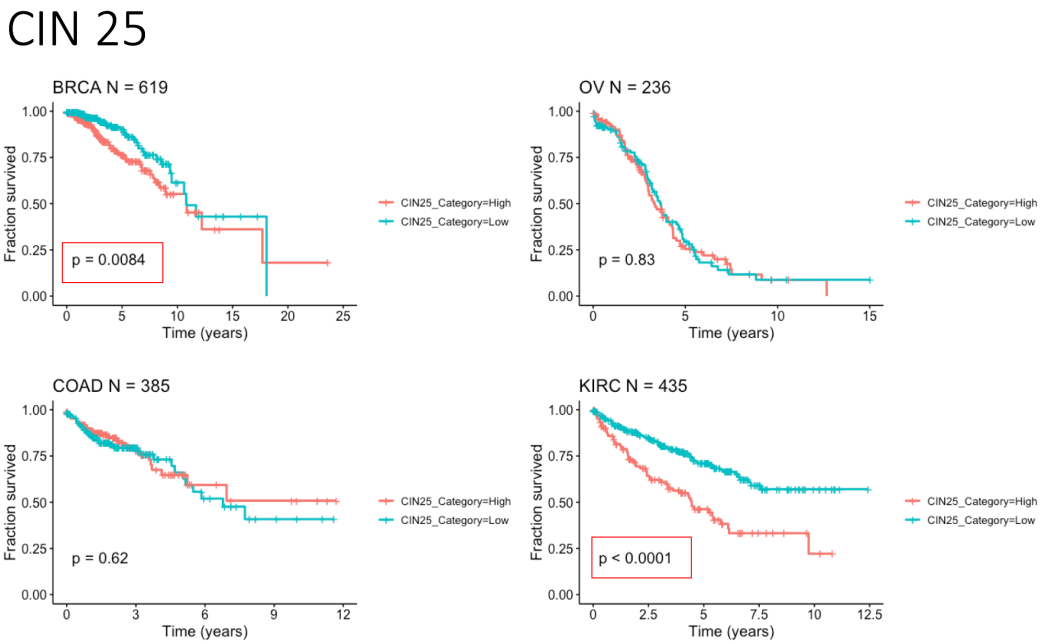


Figure 6. Univariate overall survival analysis for CIN25 scores across BRCA, OV, COAD, and KIRC tumor types. The BRCA and KIRC datasets had significant results with p-values = 0.0084, and <0.0001, respectively.

For the overall survival using the GI score, the results for the BRCA and KIRC datasets were significant and consistent, where a lower score is associated with better outcomes (Figure 7). The OV dataset was also significant, and consistent with the findings from the Zhang et. al. paper [11], where a lower score is associated with worse outcomes. The COAD datasets was not significant. The OV dataset was the only tumor type assessed in the Zhang et. al. paper [11], indicating that the outcomes are specific to tumor type. The GI score can be used to predict the outcomes for the BRCA and KIRC dataset, however the outcomes are the opposite for the OV dataset. Again, I will assess the multivariate analysis to identify any cofounding factors that could possible affect the outcome prediction.

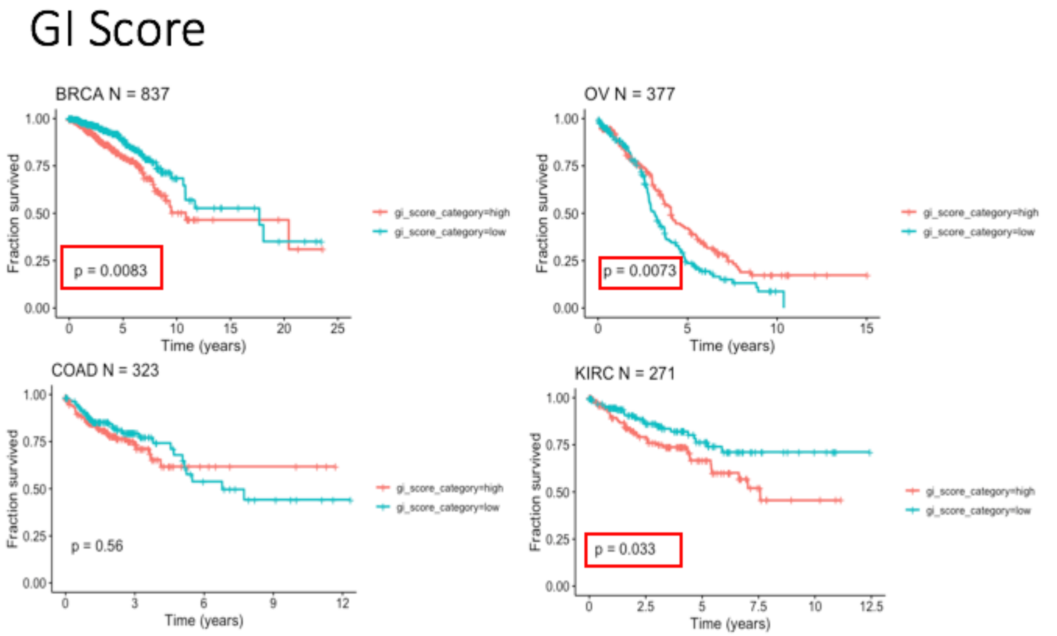


Figure 7. Univariate overall survival analysis for GI scores across BRCA, OV, COAD, and KIRC tumor types. The BRCA, OV, and KIRC datasets had significant results with p-values = 0.0083, 0.0073, and 0.033, respectively.

The overall survival stratified by tumor stage, using the CIN25 score, revealed that scores were predictive of BRCA and KIRC late stage tumors (Figure 8). The early stage tumors were not significant. The outcomes were consistent where a lower score is associated with better outcomes. This indicates that CIN25 scores can be used to predict patient outcomes for BRCA, and KIRC datasets in late stage tumors, but not for early stage.

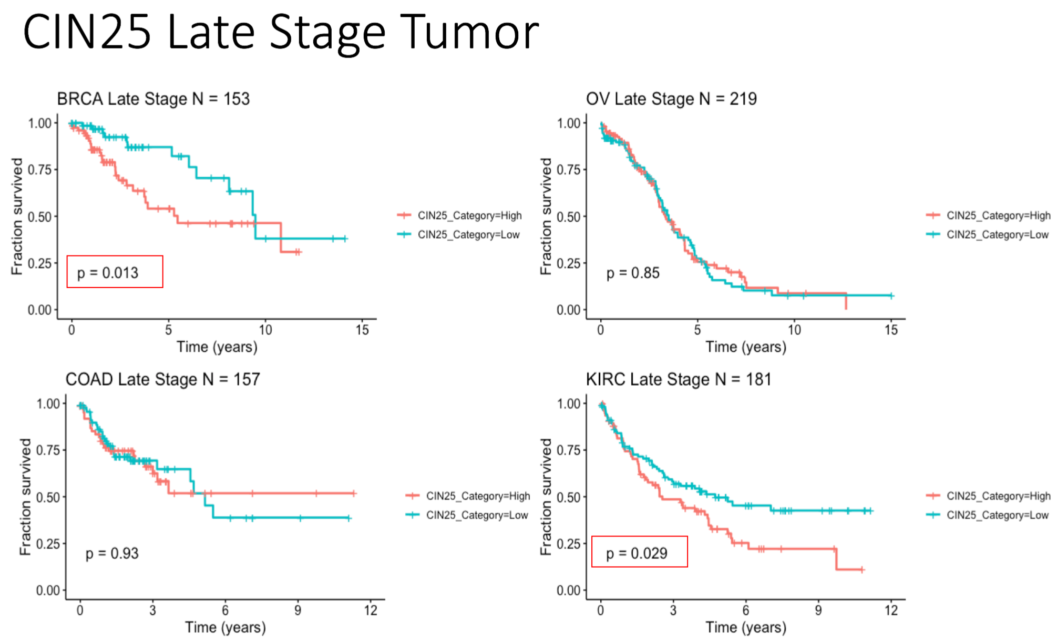
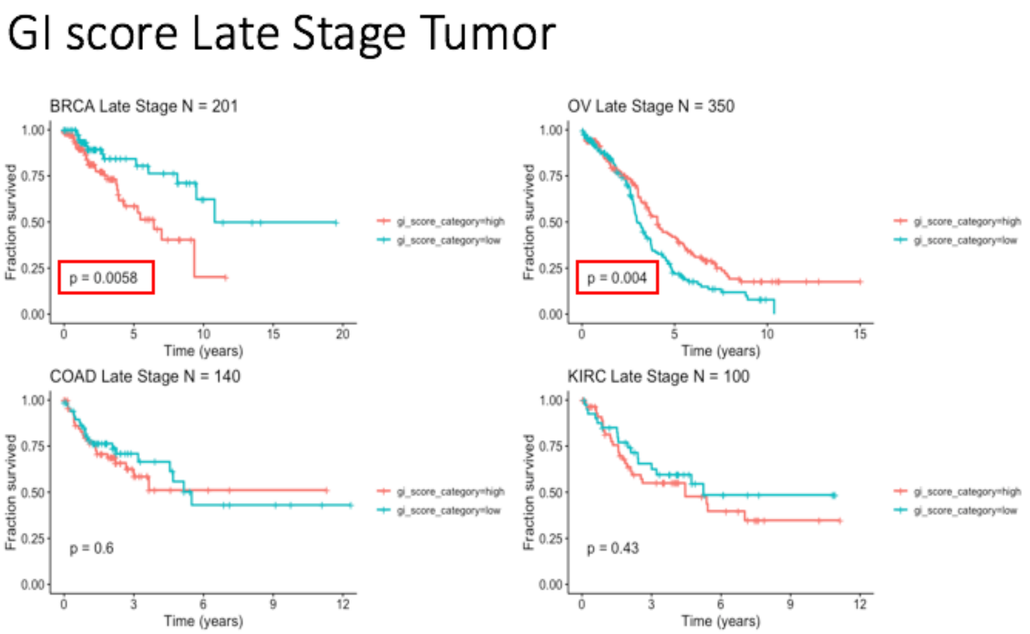

CIN25 Late Stage Tumor

Figure 8. Univariate overall survival analysis for CIN25 scores across late stage BRCA, OV, COAD, and KIRC tumor types. The BRCA and KIRC datasets had significant results with p-values = 0.013 and 0.029, respectively.

The overall survival stratified by tumor stage, using the GI scores, revealed that scores were predictive of BRCA and OV late stage tumors (Figure 9). The early stage tumors were not significant. The outcomes were consistent with the GI score across all tumors, where a lower score is associated with better outcomes for the BRCA dataset and a lower score is associated with worse outcomes for the OV dataset. This indicates that GI scores can be used to predict patient outcomes for BRCA and OV datasets in late stage tumors, but not for early stage. Also, note that the GI scores for the KIRC dataset were not significant once they were stratified by tumor stage, indicating that the CIN25 is better at predicting outcomes for the KIRC dataset.



GI score Late Stage Tumor

Figure 9. Univariate overall survival analysis for GI scores across late stage BRCA, OV, COAD, and KIRC tumor types. The BRCA and OV datasets had significant results with p-values = 0.0058 and 0.004, respectively.

There were no significant survival curves when I used the recurrent abnormal regions that were in >20% and >10 subjects for the GI scores. The breast cancer data did not have enough regions to categorize high and low scores (Figure 10). However, when I used all recurrent abnormalities, the BRCA and OV datasets were significant and reflected the same results as the full GI score (Figure 11). The lower scores were associated with better survival in the BRCA dataset, whereas lower scores were associated with worse survival in the OV data. The kidney cancer dataset was almost significant at p=0.051, however it does show the same trend where a lower score is associated with better survival. The difference between these scores and the full GI score is that a small percentage of patients changed from low to high, or high to low scores. The modified GI score has not been implemented in previous studies, but it supports the same findings as the full GI score and is the beginning step to calculating a cancer-type specific score.

GI Score (recurrent cases > 20% and 10)

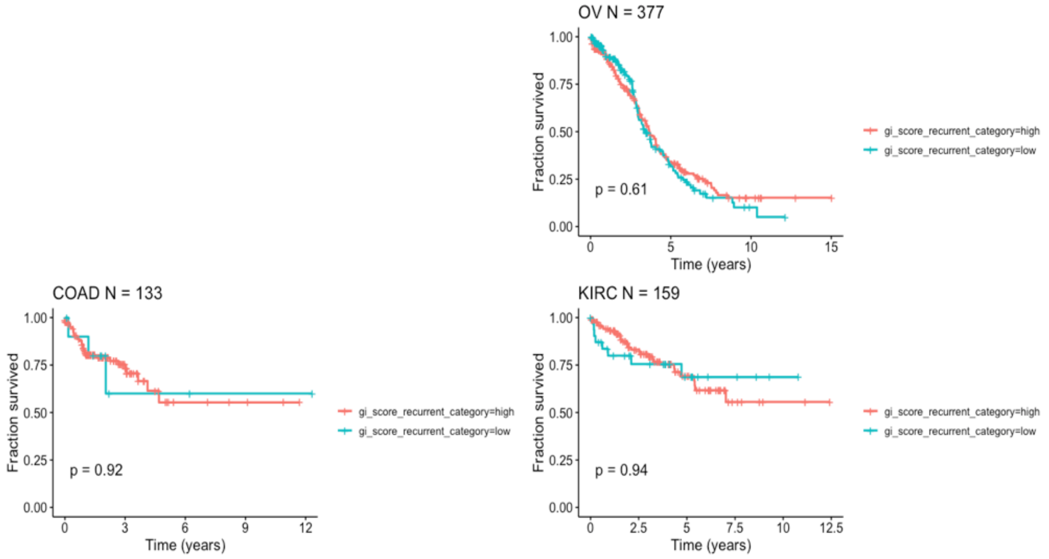


Figure 10. Univariate overall survival analysis for modified GI scores across OV, COAD, and KIRC tumor types. None of these were significant. The modified GI score includes recurrent abnormality regions in >20%, and 10 subjects.

GI Score (all recurrent cases)

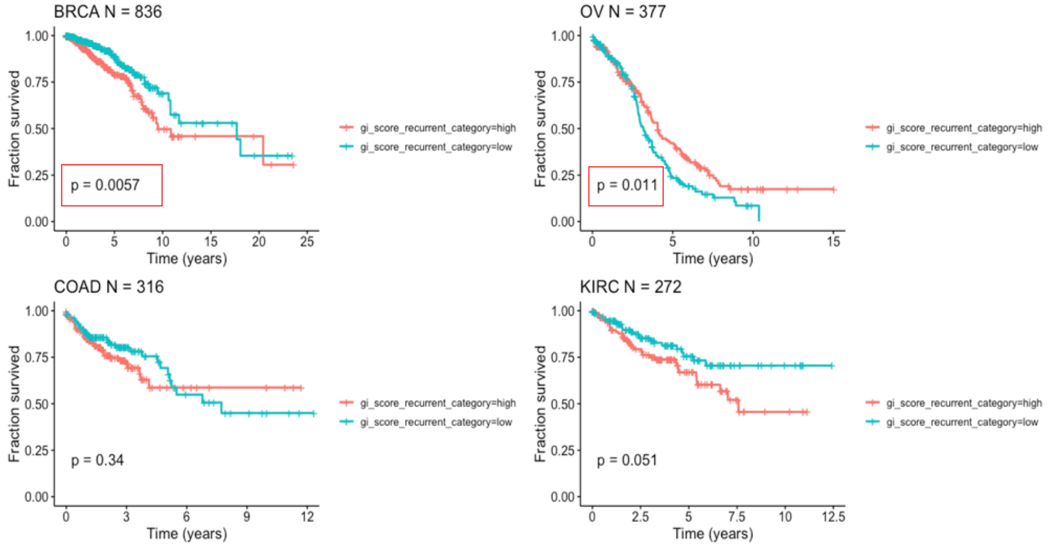


Figure 11. Univariate overall survival analysis for modified GI scores across BRCA, OV, COAD, and KIRC tumor types. The BRCA and OV datasets were significant at the p< 0.05 level.

For the progression-free survival, the BRCA dataset was significant for both the CIN25 and GI score (Appendix 2.1-2.2). The sample sizes were much lower for the other three tumor types, indicating a lack of statistical power to detect outcomes. However, the findings were consistent, where the lower score predicted better outcomes for progression-free survival, indicating that patients with lower CIN25 and GI scores are less likely to have recurrent tumors than patients with higher CIN25 and GI scores. The score, again, becomes irrelevant after about 10 years.

3.2.a Multivariate analysis

For the multivariate analysis, I computed the base model using age at dx, gender, and tumor stage to identify if any of these were significant to survival outcomes. The reason for including these covariates in the model is that they are common survival predictors that could be confounding with the scores. Also, there was a significant difference in age at dx stratified by CIN25 score for breast cancer patients, age at dx stratified by GI score for kidney cancer patients. Higher CIN25 scores were skewed in younger breast cancer patients, and higher GI score was skewed in older kidney cancer patients. In addition, there was a significant difference in tumor stage stratified by scores for kidney cancer patients, where low scores were found in early stage tumors. To the base model, I added the CIN25 category and the GI score category both separately and together with the base covariates to identify if the scores provide different information when combined with other clinical features. I computed these models for both overall and progression-free survival.

For the overall survival base model, both age at dx and tumor stage were significant to patient outcomes, where older age and later stage were associated with worse outcomes (Appendix 2.3-2.6). After adding the CIN25 category to this model, it was only significant for the BRCA and KIRC datasets, where the findings were consistent in that low score were associated with better outcomes (Appendix 2.7-2.10). Furthermore, after adding both the CIN25 category and the GI category to the base model, the only score that was significant was the CIN25 score for the KIRC dataset (Appendix 2.11-2.14). I also computed the overall survival model with the base covariates and the GI category, where the GI category was only significant for the BRCA dataset (Appendix 2.15-2.18). Lastly, I computed the multivariate analysis for late stage ovarian tumors. This revealed that the ovarian late stage model had a significant finding that was consistent with the univariate analysis, further supporting that late ovarian cancer can be differentiated by GI scores, where a lower GI score is associated with worse prognosis (Appendix 2.19).

For the progression-free survival base model, the tumor stage was significant to patient outcomes, where later stage was associated with worse outcomes (Appendix 2.20-2.22). Being male was associated with worse outcomes in the KIRC dataset (HR=3.9, p=0.014). After adding the CIN25 category to the model, tumor stage was significant was significant for all tumor types, and gender was significant for the COAD and KIRC datasets, where being male was associated with worse outcomes (Appendix 2.23-2.25). Adding the GI score to this model did not add any information and the same covariates were significant; tumor stage was significant for all tumor types, and gender was significant for the COAD and KIRC datasets. Similarly, the base model with the GI category did not add any additional information and had the same findings as the previous models.

Chapter 4: Discussion

4.1 Overview of findings

These findings show that chromosomal instability is, in fact, specific to tumor type and that the clinical outcomes are not consistent. Overall, univariate survival analysis revealed that CIN25 scores were predictive of breast and kidney cancer, where lower scores were associated with better outcomes. Similarly, the GI score was predictive of breast, kidney and ovarian cancer. However, ovarian cancer GI scores show a negative association with outcome, where a lower score was associated with worse outcomes. The scores were not significantly predictive of the colon cancer dataset. In addition, the modified GI scores that contained recurrent abnormality reflected the same trends as seen from the full GI score. From both the univariate and multivariate analyses, it can be concluded that the chromosomal instability scores provide different information for diverse tumor types.

Even though there were less late stage tumors with low CIN25 scores in the kidney cancer dataset, the CIN25 score was still able to predict outcomes for late stage tumors. Similarly, the CIN25 and GI score predicted outcomes for late stage breast cancer tumors, where lower scores were associated with better outcomes. Also, the GI score was able to predict outcomes for late stage ovarian cancer, where a lower score was associated with worse prognosis. This indicates that patients diagnosed at later stages can be further stratified based on their CIN25 or GI score. Since it is known that later stages have poorer survival, the clinical implications give an opportunity to select patients that may respond to treatment or predict prognosis based on instability scores.

Progression-free univariate analysis revealed that the breast cancer dataset was the only tumor type that had significant outcomes, where lower score was associated with better outcomes. However, sample sizes for this analysis were much lower compared to the overall survival because of missing data across samples. Because the same trend occurred for the breast cancer dataset, and the breast cancer dataset had more samples than the other tumor types, it is possible the progression-free survival analysis reflects a lack of statistical power to detect a difference in scores.

Age at diagnosis and tumor stage were significant predictors in the overall survival multivariate analyses, however, the GI score was significant for the breast cancer dataset when added to the base model. In addition, the CIN25 was significant for both breast and kidney when added to the base model. This indicates that the GI score and CIN25 were predictive for breast cancer outcomes, even if they were confounding with age at diagnosis and tumor stage. However, when both scores are included in the same model, neither score was significant. It is possible the scores were not predictive of breast cancer when combined because these scores were the most correlated out of all the tumor types at 0.5 correlation (Table 1). In contrast, the CIN25 score was significant for kidney cancer when combined with the GI score, and the GI score was not significant. This indicates that the CIN25 is a better predictor for kidney cancer outcomes.

None of the scores were significant in the multivariate analysis of progression-free survival. Tumor stage was a significant covariate for the BRCA, KIRC, and OV datasets. In addition, gender was significant for the colon and kidney cancer datasets, where being male was associated with worse outcomes. Age was not significant, in contrast to the overall multivariate analysis. This indicates that although, progression may not be predicted from scores, staging information can be further stratified using scores.

4.2 Recommendations

Because clinical decisions are based on many factors, the primary goal of the chromosomal instability scores is to supplement, rather than replace known covariates, like tumor stage, age, and gender. Although these known covariates can predict outcomes in many cancer types, the chromosomal instability scores provide additional information to further stratify tumors. The scores indicate that a chromosomal instability trend exists upon breast and kidney tumors that we do not see in colon cancers. Also, a trend exists among ovarian cancer that is opposite of breast and kidney cancers. Given these trends, I would recommend using the CIN25 score to differentiate kidney tumors, and either GI scores or CIN25 scores, to differentiate breast tumors. Late stage ovarian cancers can be stratified using GI scores, where a higher score is associated with worse outcomes.

4.3 Significance to the field

There were a total of 12 tumor type and score comparisons, where the majority of these contribute new findings to the field of genomic instability. The breast and ovarian tumor types were used in the previous CIN25 study and the ovarian tumor type was assessed in the GI study, however these datasets did not use the most recent TCGA data and the GI study did not take into account particular patient cohorts such as last stage. Therefore, these findings contribute 10 new findings to the field with regard to tumor type and score comparisons.

The contributions of this thesis point to a need for cancer-type specific genomic instability scores. This is because no single score predicted all cancer types equally well. This means that cancer-type specific genomic instability scores need to be done on a case by case basis. The GI recurrent score is the beginning step in developing cancer-type specific scores. It could be further developed beyond what is shown here, which I will discuss more in the future directions, section 5.2. These new findings also highlight the fact cancer-type specific scores could be used for particular patient subsets such as late-stage tumors.

Chapter 5: Conclusions

5.1 Summary

In summary, the CIN25 score, GI score, and modified GI score have provided different information across tumor types in both the univariate and multivariate survival analyses. Specifically, the CIN25 predicted that lower scores were associated with better patient outcomes in kidney cancer, the GI and CIN25 score predicted that lower scores were associated with better patient outcomes in breast cancer, the GI score predicted that lower scores were associated with worse patient outcomes in late stage ovarian tumors, and none of these scores predicted the outcomes of colon cancer.

The contributions of this thesis show that there is a need for a cancer-type specific genomic instability score based on the results of the predictive power of genomic instability scores and how they differ across tumor types. In addition, the results of this project contribute to the development of characterizing chromosomal instability of solid tumors, and can potentially help improve patient care by creating the beginning step in a cancer-type specific genomic instability score that utilizes both structural abnormalities as well as genomic aberrations, which can further the understanding of how these scores can stratify specific tumor types.

5.2 Limitations

The sample sizes were not consistent across tumor types because it was possible that some samples had genomic data, but not clinical data, and vice versa. Also, within the clinical data, there missing values that could not be included in the survival analysis. The lack of consistent sample sizes across data types contributed to the reduction in statistical power for the progression-free survival analysis. The Mitelman database also contained small sample sizes of recurrent abnormalities.

5.3 Future work

These findings have the potential for new areas of therapy development, refining clinical decision making, or correlating scores with response to treatment in solid tumors. For example, the kidney cancer is the one tumor type that has not been implemented in the previous studies, and the scores consistently predicted overall survival outcomes even after adjusting for stage. Future studies could assess if there is an association between treatment outcomes and GI score. In addition, the fact that the scores did not perform well for the colon cancer dataset indicate that the scores could potentially be modified to stratify colon tumors more effectively. For example, there might be a more relevant gene signature related to colon cancer.

The ovarian dataset was the deadliest cancer out of all four tumor types because there were significantly more deceased patients, progressors, and late stage tumors than any other cancer type. Therefore, there is a significant opportunity to further develop treatment for ovarian cancer. This is particularly true for patients with high GI scores since they have significantly less mutations, less altered copy number regions, and worse prognosis than late stage ovarian patients with higher GI scores. Personalized medicine for ovarian cancer could be improved by stratifying late stage tumors based on score.

Many studies have shown that ovarian tumors with both BRCA1/2 mutations is associated with better prognosis, whereas BRCA1 in breast cancer is associated with worse prognosis [26]. This might be contributing to the fact that the survival outcomes were the opposite for these tumor types. Overexpression of HER2 is another genomic feature that is frequently used to differentiate breast cancer tumors [27]. Therefore, it might be more informative to adjust the scores, or further stratify tumors, based on BRCA1/2 and HER2 when predicting the outcomes of breast and ovarian cancer datasets.

The GI score has potential to be quantified in other ways by utilizing mutation only scores, or copy number only scores, and the weight could be adjusted from the default of 0.5. The mutation only score could be the number of mutations for each patient, or the number of mutations times a weighted value. Similarly, the copy number could be the number of copy number regions or the number of copy number regions times a weighted value.

For the modified GI score that involved recurrent abnormalities, a negative control could be developed that would include permutations of sampling the same number of genomic regions of copy number and mutation that contained recurrent abnormalities. It could also be used to identify which cytobands that are most predictive to develop a cancer type-specific signature.

Another way recurrent abnormalities could be integrated into the GI score is by utilizing a structure-specific recurrent abnormality, such as regions that only exist on isochromosomes. In addition, the overlap of recurrent abnormalities across tumor types could be assessed for each tumor type to see if these predict clinical outcomes. Finally, as more recurrent structural abnormalities are identified, these can potentially be incorporated into the GI scores to predict outcomes more precisely than clinical measures.

References

1. Yao Y, Dai W. Genomic instability and cancer. J Carcinog Mutagen. 2014;5(2):1-5.
2. Nanjangud G, Amarillo I, Rao PN. Solid tumor cytogenetics: current perspectives. Clin Lab Med. 2011;31(4):785-811, xi.
3. National Human Genome Research Institute [Internet]. Bethesda, MD: NHGRI; 2003-2018. Chromosome abnormalities; 6 January 2016 [cited 2018 Nov 12]. Available from: https://www.genome.gov/11508982/chromosome-abnormalities-fact-sheet/
4. Wan TSK. Cancer Cytogenetics Methods and Protocols. New York, NY: Humana Press; 2017.
5. Das K, Tan P. Molecular cytogenetics: recent developments and applications in cancer. Clin Genet. 2013;84(4):315-25.
6. Sudoyo AW, Hardi F. Cytogenetics in solid tumors: lessons from the Philadelphia Chromosome. Acta Med Indones. 2011;43(1):68-73.
7. Ley TJ, Miller C, Ding L, Raphael BJ, Mungall AJ, Robertson A, et al. Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. N Engl J Med. 2013;368(22):2059-74.
8. Varella-Garcia M. Molecular cytogenetics in solid tumors: laboratorial tool for diagnosis, prognosis, and therapy. Oncologist. 2003;8(1):45-58.
9. Geigl JB, Obenauf AC, Schwarzbraun T, Speicher MR. Defining 'chromosomal instability'. Trends Genet. 2008;24(2):64-9.

10. Carter SL, Eklund AC, Kohane IS, Harris LN, Szallasi Z. A signature of chromosomal instability inferred from gene expression profiles predicts clinical outcome in multiple human cancers. Nat Genet. 2006;38(9):1043-8.

11. Zhang S, Yuan Y, Hao D. A genomic instability score in discriminating nonequivalent outcomes of BRCA1/2 mutations and in predicting outcomes of ovarian cancer treated with platinum-based chemotherapy. PloS One. 2014;9(12):e113169.

12. Song L, Bhuvaneshwar K, Wang Y, Feng Y, Shih IM, Madhavan S, et al. CINdex: A Bioconductor Package for Analysis of Chromosome Instability in DNA Copy Number Data. Cancer Inform. 2017;16.

13. Grossman RL, Heath AP, Ferretti V, Varmus HE, Lowy DR, Kibbe WA, et al. Toward a shared vision for cancer genomic data. N Engl J Med. 2016;375(12):1109-12.

14. Seshan VE, Olshen A (2018). *DNAcopy: DNA copy number data analysis*. R package version 1.56.0.

15. Cibulskis, Kristian, Michael S. Lawrence, Scott L. Carter, Andrey Sivachenko, David Jaffe, Carrie Sougnez, Stacey Gabriel, Matthew Meyerson, Eric S. Lander, and Gad Getz. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat Biotech. 2013;31(3): 213-219.

16. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GR, Thormann A, et al. The Ensembl variant effect predictor. Gen Bio. 2016;17(1):122.

17. Mitelman F, Johansson B and Mertens F. Mitelman database of chromosome Aberrations and gene fusions in cancer [Internet]. Bethesda, MD: NCI; 2018 [cited 2018 Nov 12]. Available from: http://cgap.nci.nih.gov/Chromosomes/Mitelman

18. Durinck S, Spellman P, Birney E, Huber W (2009). Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat Protoc*. 2009; 4(8):1184–1191.

19. Durinck S, Moreau Y, Kasprzyk A, Davis S, De Moor B, Brazma A, Huber W. BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinform*. 2005;21(16):3439–3440.

20. Yin T, Lawrence M, Cook D (2018). *biovizBase: Basic graphic utilities for visualization of genomic data.* R package version 1.30.0.

21. Sotiriou C, Wirapati P, Loi S, Harris A et al. Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. J Natl Cancer Inst. 2006 Feb 15;98(4):262-72.

22. Carvalho BS, Irizarry RA. A framework for oligonucleotide microarray preprocessing. Bioinform. 2010;26(19):2363-7.

23. Gentleman R, Carey V, Huber W, Hahne F (2018). *genefilter: genefilter: methods for filtering genes from high-throughput experiments*. R package version 1.64.0.

24. Pearson, K. Notes on regression and inheritance in the case of two parents. Proc R Soc Lond. 20 June 1895;58:240–242.

25. Therneau TM, Grambsch PM. *Modeling Survival Data: Extending the Cox Model.* New York: Springer; 2000.

26. Zhong Q, Peng HL, Zhao X, Zhang L, Hwang WT. Effects of BRCA1- and BRCA2-related mutations on ovarian and breast cancer survival: a meta-analysis. Clin Cancer Res. 2015;21(1):211-20.

27. Ahmed S, Sami A, Xiang J. HER2-directed therapy: current treatment options for HER2-positive breast cancer. Breast Cancer (Tokyo, Japan). 2015;22(2):101-16.

Appendix

Appendix 1: EDA Results

Appendix 2: Survival Analyses

Appendix 1: EDA Results

Appendix 1 provides all the EDA plots for the scores including the correlations between scores, score distributions across tumor types, and scores stratified by clinical features.

Figure A1.1. Correlation plots of scores across tumor types.

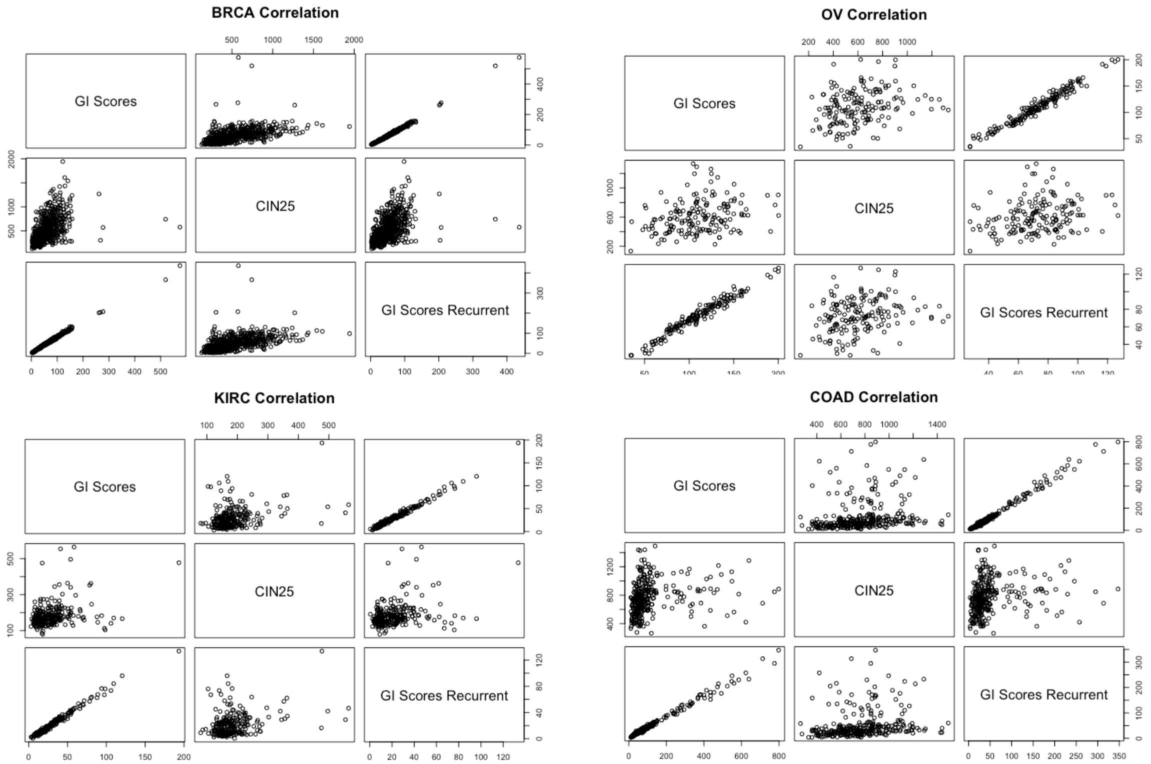


Figure A1.2. Distributions of raw scores across tumor types.

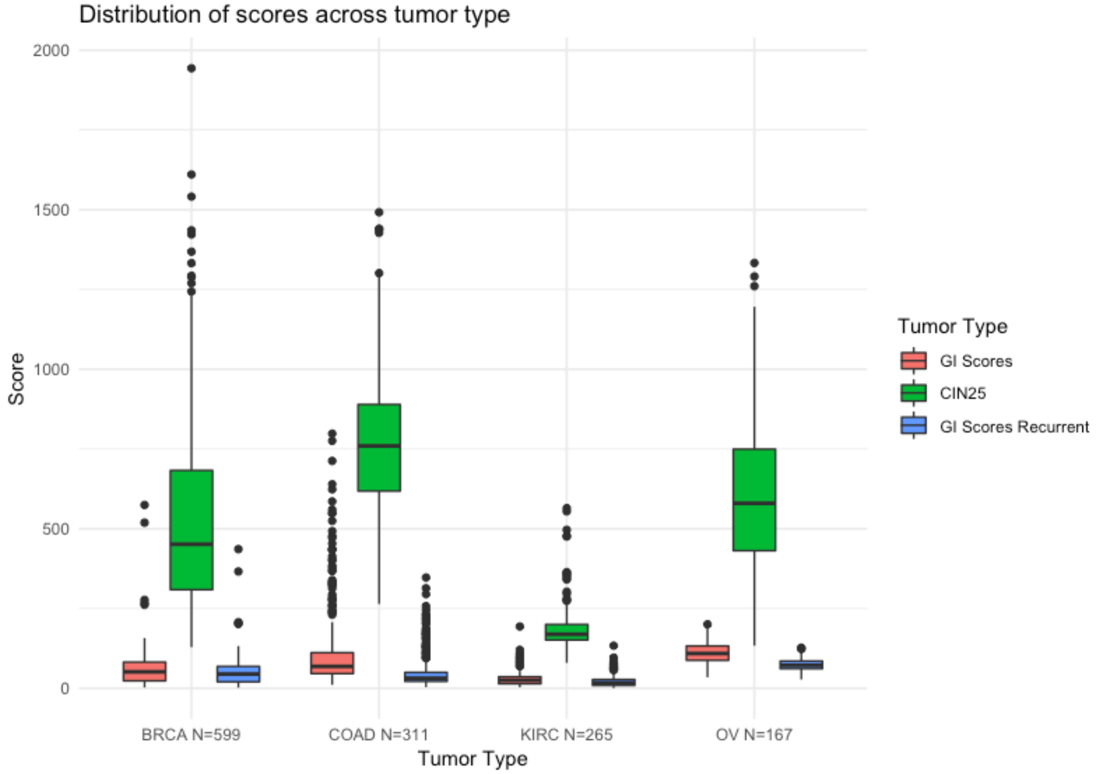


Table A1.3. ANOVA of standardized scores showing a significant difference in scores across tumor types at the p<0.05 level.

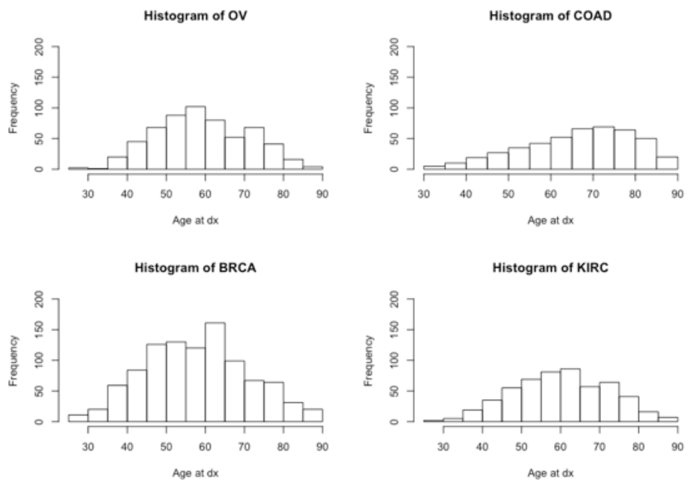
	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Tumor Type	3	1201	400.4	568.6	<2e-16*
Score Type	2	0	0	0	1
Residuals	4050	2852	0.7		

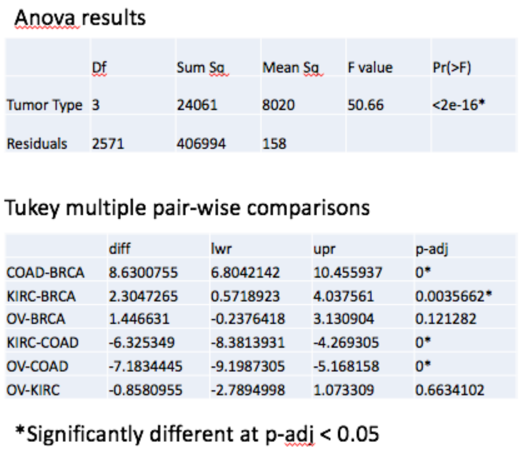
Table A1.4. Tukey multiple pair-wise comparisons showing significant differences in standardized scores between each tumor type comparison. The diff column shows the difference in the means between the two tumor types, and the lwr and upr columns show the end point of the 95% confidence interval. However, there were no significant differences across score types.

	diff	lwr	upr	p-adj
COAD- BRCA	0.7445381	0.65832	0.8307562	0*
KIRC-BRCA	-0.8609728	-0.9527179	-0.7692277	0*
OV-BRCA	0.4282014	0.3192402	0.5371627	0*
KIRC-COAD	-1.6055109	-1.7088256	-1.5021962	0*
OV-COAD	-0.3163367	-0.4352034	-0.1974699	0*
OV-KIRC	1.2891742	1.1662397	1.4121088	0*

	diff	lwr	upr	p-adj
CIN25_z-gi_scores_z	-3.20E-16	-0.07567088	0.07567088	1
CIN70_z-gi_scores_z	-3.16E-16	-0.07567088	0.07567088	1
CIN70_z-CIN25_z	3.79E-18	-0.07567088	0.07567088	1

Figure A1.5. Overall comparison of distributions of age at dx across tumor types.





Anova results

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Tumor Type	3	24061	8020	50.66	<2e-16*
Residuals	2571	406994	158		

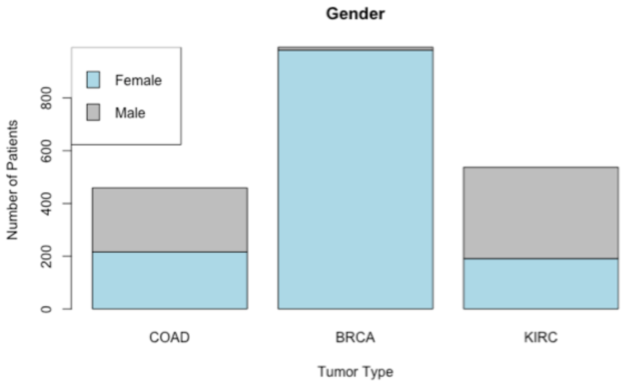
Tukey multiple pair-wise comparisons

	diff	lwr	upr	p-adj
COAD-BRCA	8.6300755	6.8042142	10.455937	0*
KIRC-BRCA	2.3047265	0.5718923	4.037561	0.0035662*
OV-BRCA	1.446631	-0.2376418	3.130904	0.121282
KIRC-COAD	-6.325349	-8.3813931	-4.269305	0*
OV-COAD	-7.1834445	-9.1987305	-5.168158	0*
OV-KIRC	-0.8580955	-2.7894998	1.073309	0.6634102

*Significantly different at p-adj < 0.05

Figure A1.6. Overall proportion of gender across tumor types.

Overall Gender



Pearson's Chi-squared test p-value: <2.2e-16*

Proportions

	COAD	BRCA	KIRC
FEMALE	216	981	191
MALE	243	11	346

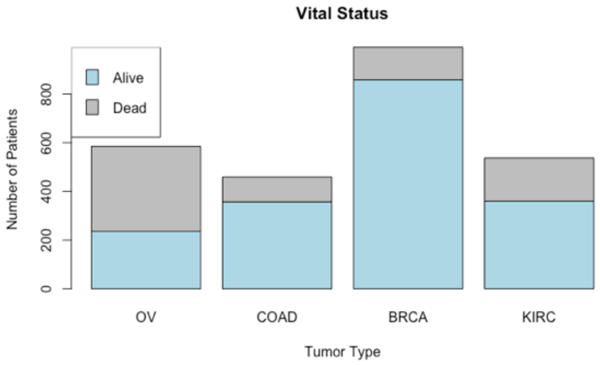
Observed/Expected

	COAD	BRCA	KIRC
FEMALE	34.05552	120.08668	90.22903
MALE	78.78178	277.80052**	208.72982

*Significant difference at 0.05 level
**Lower proportion of male patients with BRCA

Figure A1.7. Overall proportion of vital status across tumor types.

Overall vital status



Pearson's Chi-squared test p-value: <2.2e-16*

Proportions

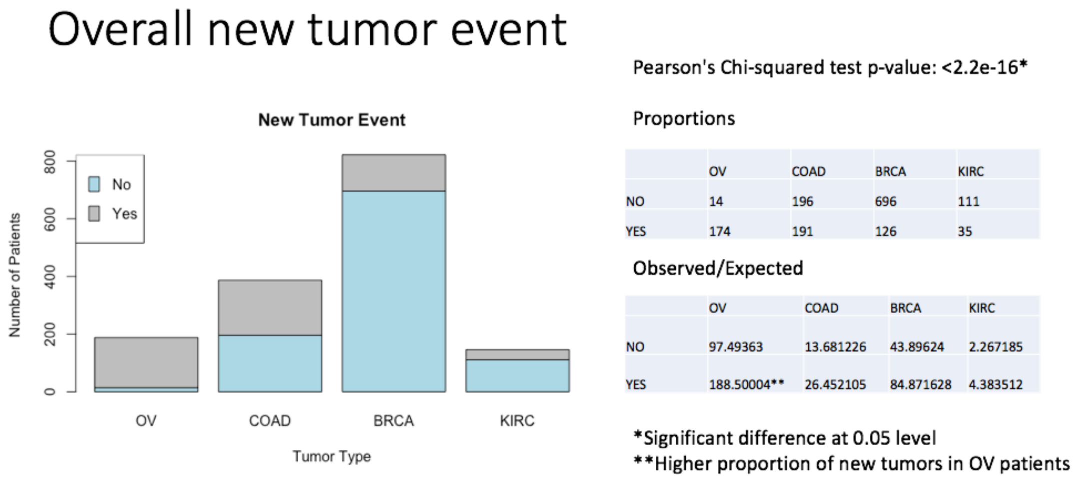
	OV	COAD	BRCA	KIRC
Alive	236	357	858	360
Dead	349	102	134	177

Observed/Expected

	OV	COAD	BRCA	KIRC
Alive	75.017132	3.564322	36.565501	0.854002
Dead	178.288749**	8.471112	86.903047	2.029656

*Significant difference at 0.05 level
**Higher proportion of deceased OV patients

Figure A1.8. Overall proportion of new tumor event across tumor types.



Overall new tumor event

Pearson's Chi-squared test p-value: <2.2e-16*

Proportions

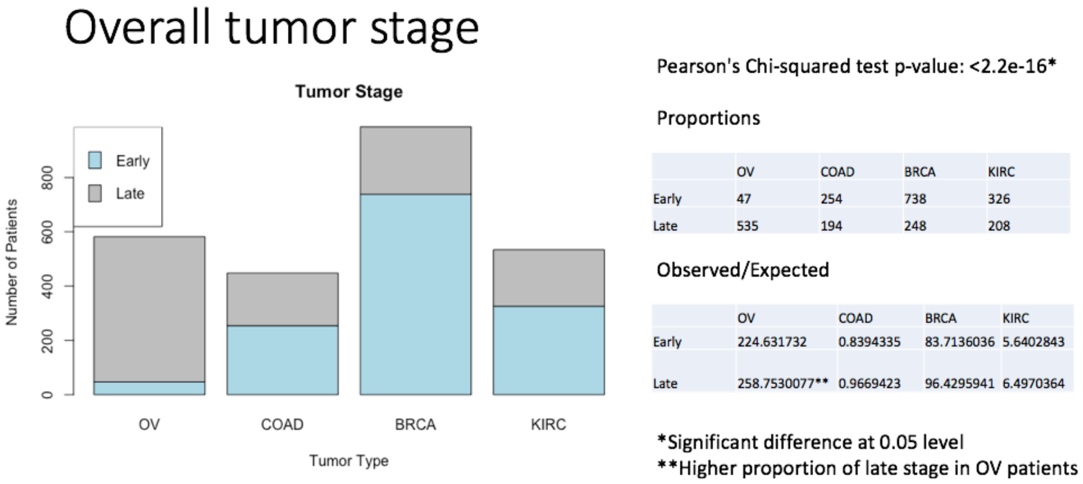
	OV	COAD	BRCA	KIRC
NO	14	196	696	111
YES	174	191	126	35

Observed/Expected

	OV	COAD	BRCA	KIRC
NO	97.49363	13.681226	43.89624	2.267185
YES	188.50004**	26.452105	84.871628	4.383512

*Significant difference at 0.05 level
**Higher proportion of new tumors in OV patients

Figure A1.9. Overall proportion of tumor stage across tumor types.



Overall tumor stage

Pearson's Chi-squared test p-value: <2.2e-16*

Proportions

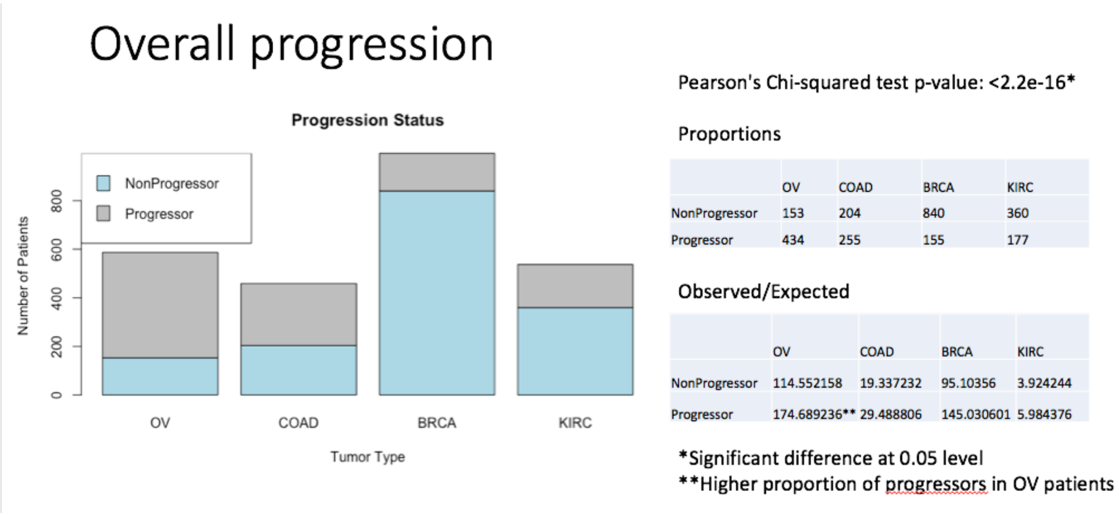
	OV	COAD	BRCA	KIRC
Early	47	254	738	326
Late	535	194	248	208

Observed/Expected

	OV	COAD	BRCA	KIRC
Early	224.631732	0.8394335	83.7136036	5.6402843
Late	258.7530077**	0.9669423	96.4295941	6.4970364

*Significant difference at 0.05 level
**Higher proportion of late stage in OV patients

Figure A1.10. Overall proportion of progression status across tumor types.



Overall progression

Pearson's Chi-squared test p-value: <2.2e-16*

Proportions

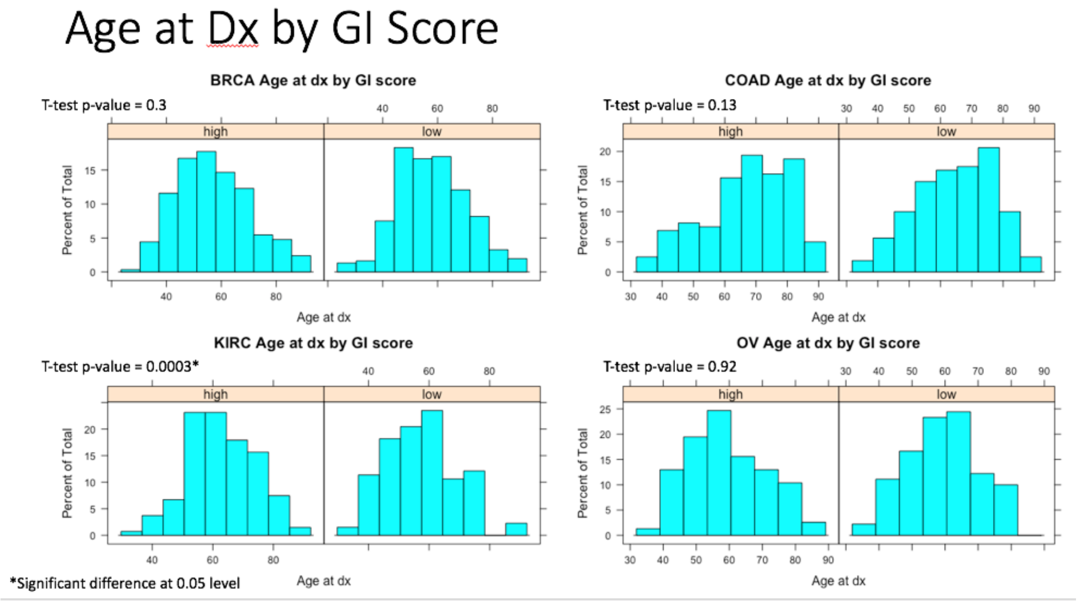
	OV	COAD	BRCA	KIRC
NonProgressor	153	204	840	360
Progressor	434	255	155	177

Observed/Expected

	OV	COAD	BRCA	KIRC
NonProgressor	114.552158	19.337232	95.10356	3.924244
Progressor	174.689236**	29.488806	145.030601	5.984376

*Significant difference at 0.05 level
**Higher proportion of progressors in OV patients

Figure A1.11. Distributions of age at dx stratified by tumor type and GI Score.



Age at Dx by GI Score

Figure A1.12. Distributions of age at dx stratified by tumor type and CIN25.

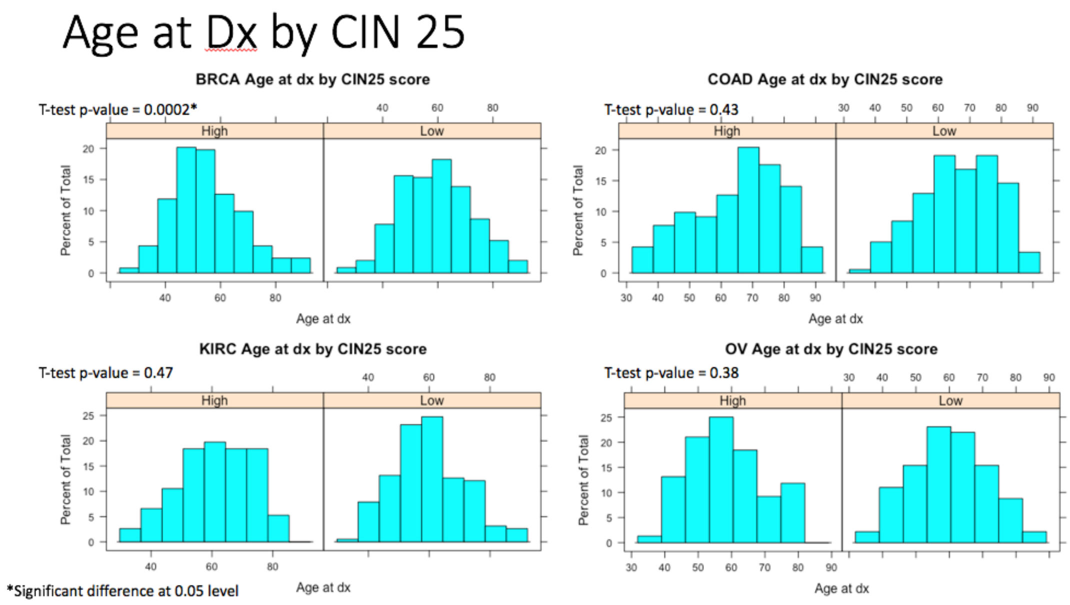


Figure A1.13. Distributions of age at dx stratified by tumor type and CIN70.

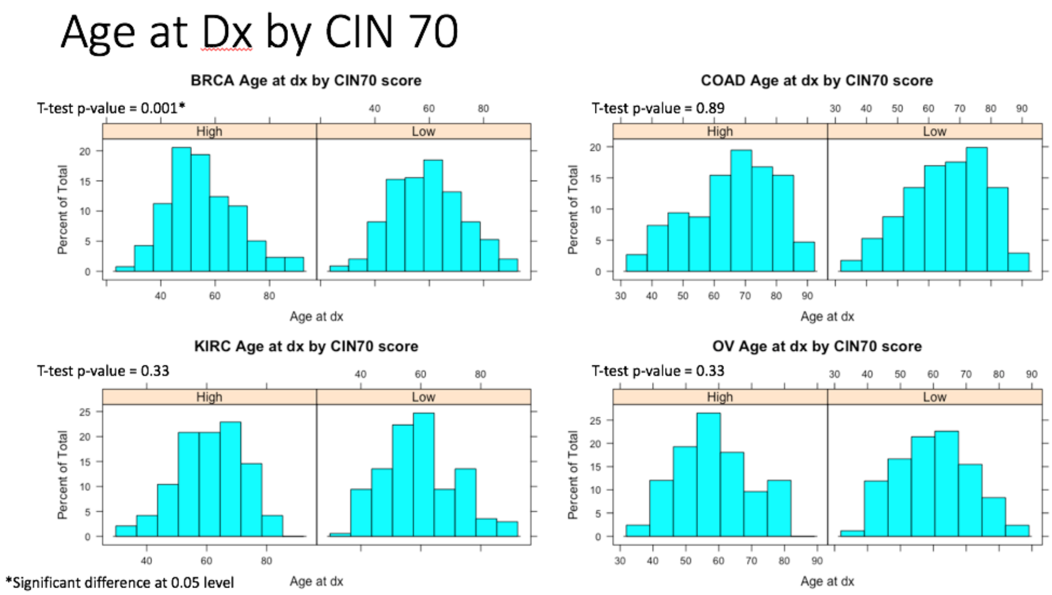
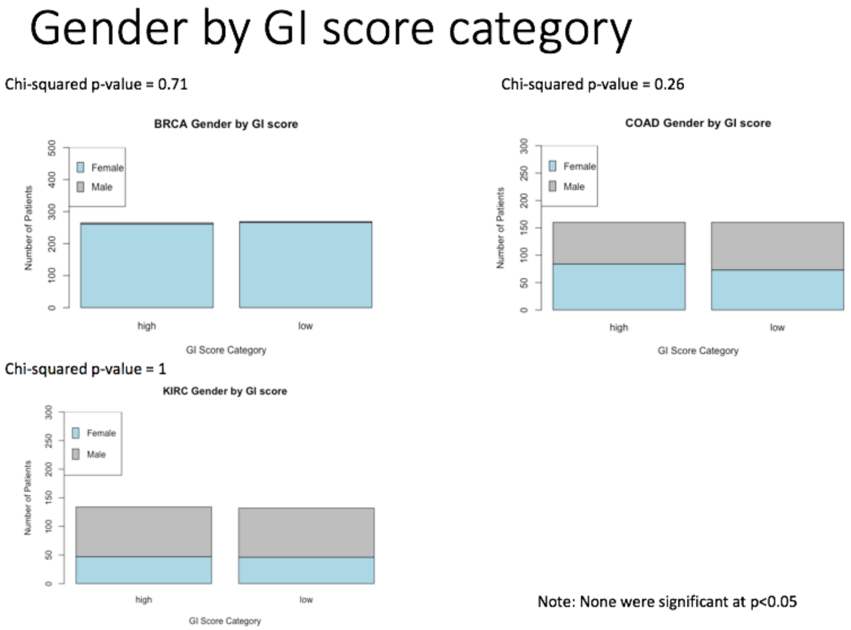


Figure A1.14. Proportions of gender stratified by tumor type and GI scores.



Gender by GI score category

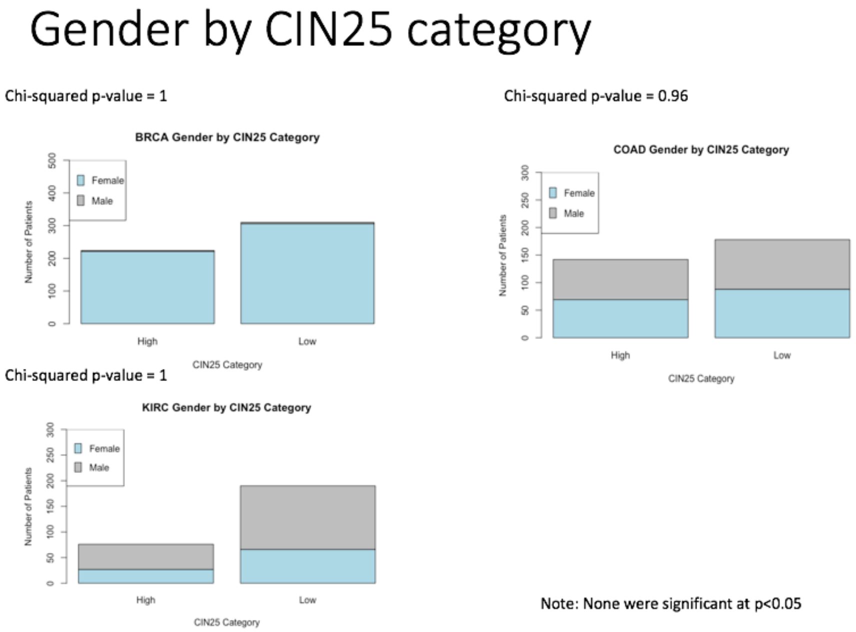
Chi-squared p-value = 0.71

Chi-squared p-value = 0.26

Chi-squared p-value = 1

Note: None were significant at p<0.05

Figure A1.15. Proportions of gender stratified by tumor type and CIN25.



Gender by CIN25 category

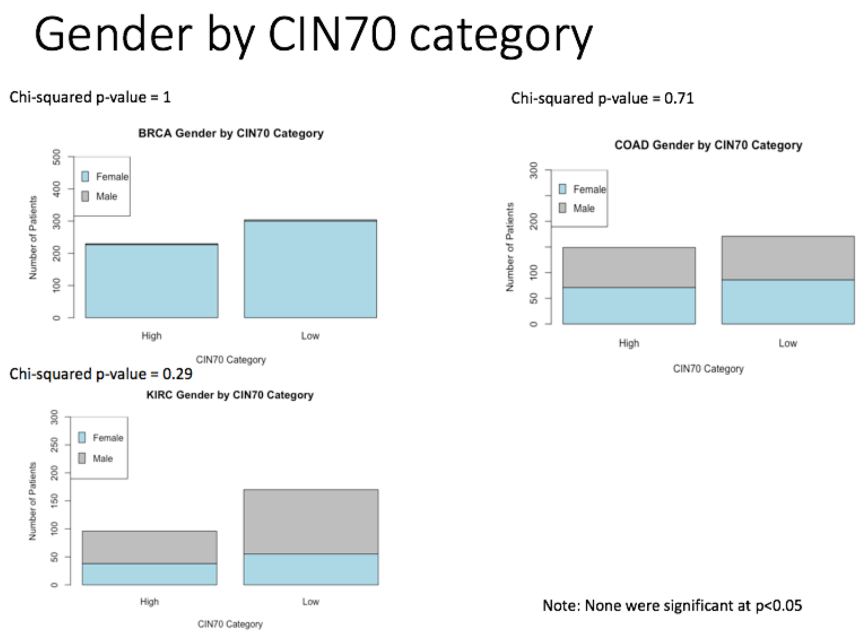
Chi-squared p-value = 1

Chi-squared p-value = 0.96

Chi-squared p-value = 1

Note: None were significant at p<0.05

Figure A1.16. Proportions of gender stratified by tumor type and CIN70.



Gender by CIN70 category

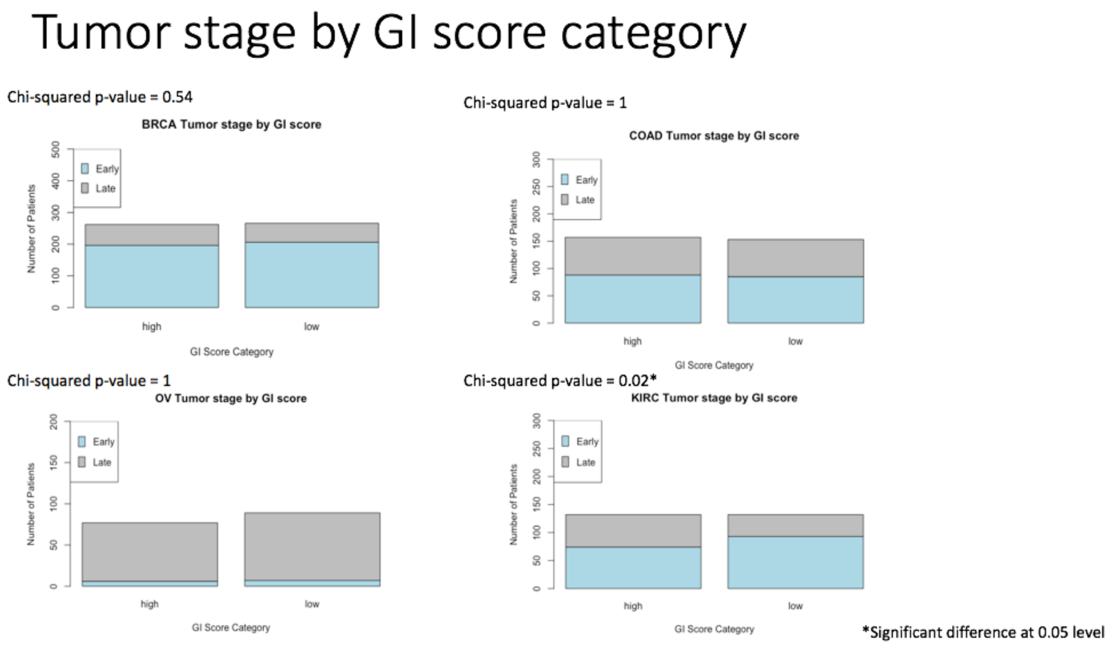
Chi-squared p-value = 1

Chi-squared p-value = 0.71

Chi-squared p-value = 0.29

Note: None were significant at p<0.05

Figure A1.17. Proportions of tumor stage stratified by tumor type and GI scores.



Tumor stage by GI score category

Chi-squared p-value = 0.54

Chi-squared p-value = 1

Chi-squared p-value = 1

Chi-squared p-value = 0.02*

*Significant difference at 0.05 level

Figure A1.18. Proportions of tumor stage stratified by tumor type and CIN25.

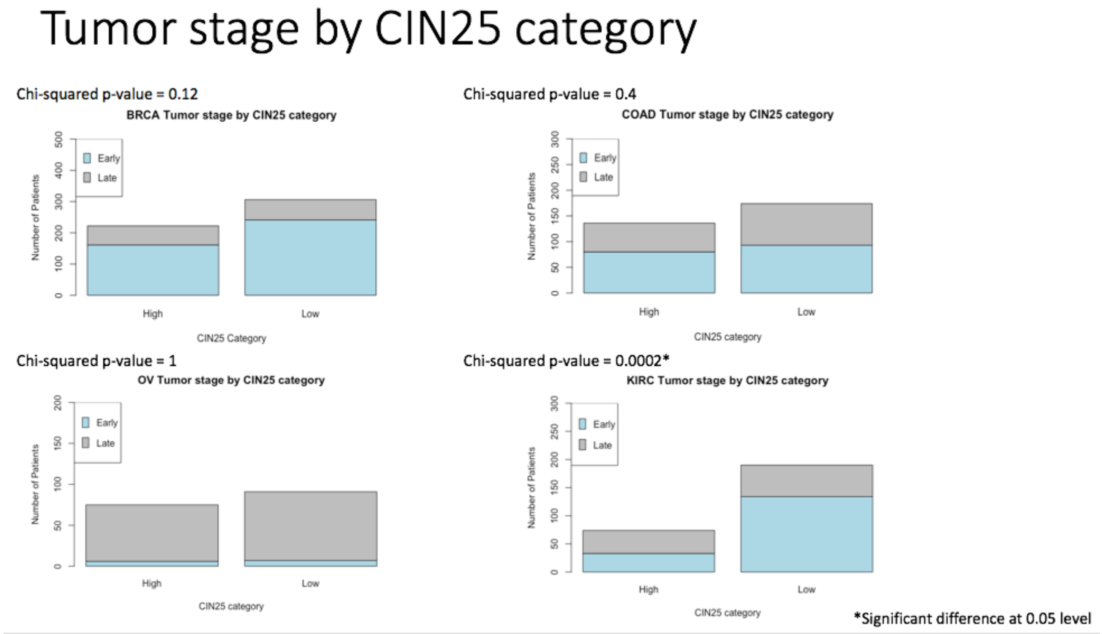
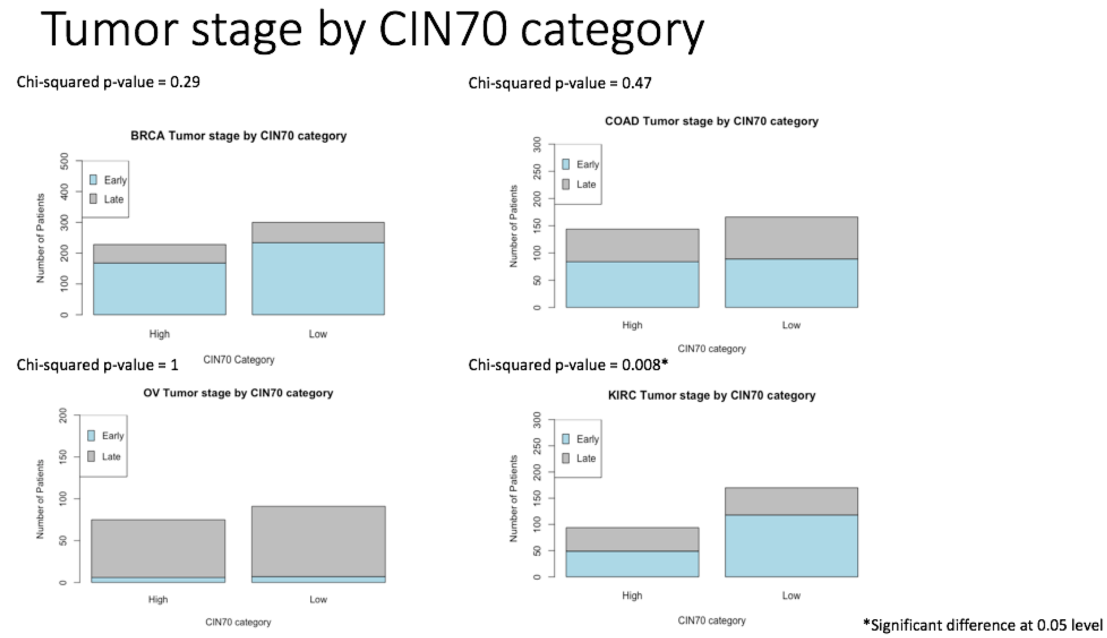


Figure A1.19. Proportions of tumor stage stratified by tumor type and CIN70.



Appendix 2: Survival Analyses

Appendix 2 provides all additional survival curves including univariate and multivariate analyses.

Figure A2.1. Univariate progression-free-survival using the CIN25 score. BRCA dataset was significant with a p=0.021.

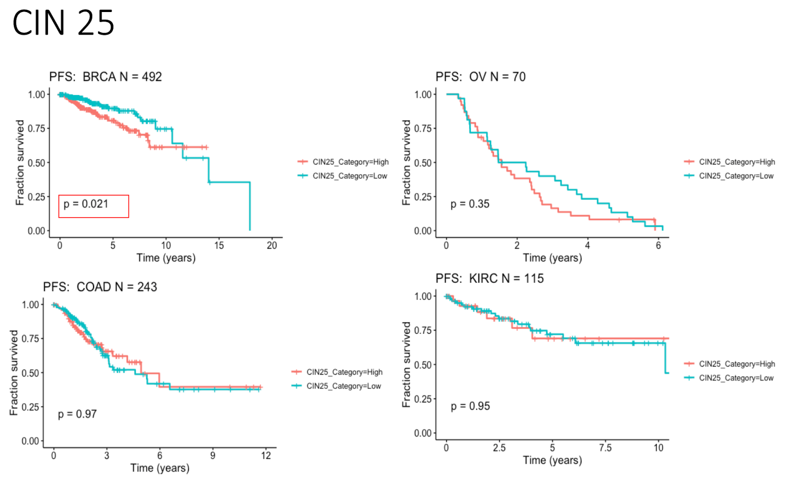
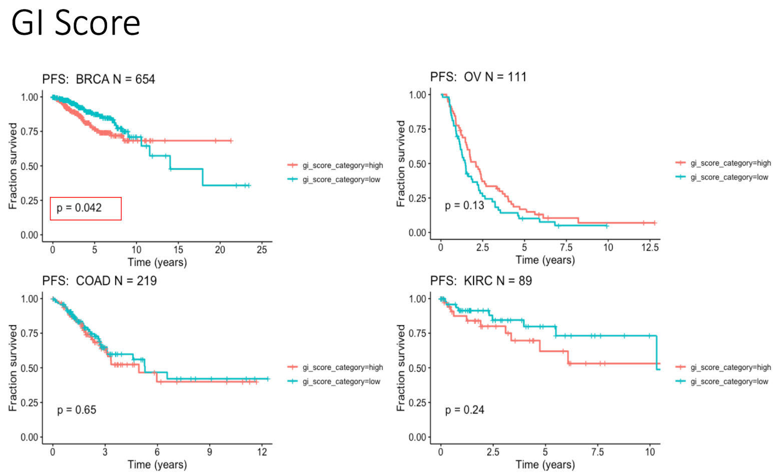
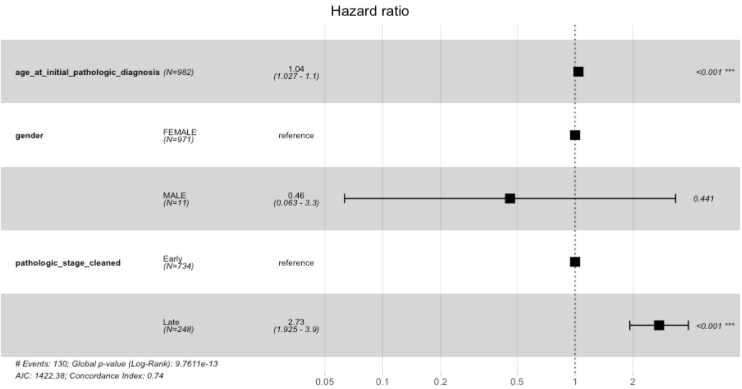


Figure A2.2. Univariate progression-free survival using the GI score. BRCA dataset was significant with a p=0.042, where a lower score is predictive of less likely to progress.



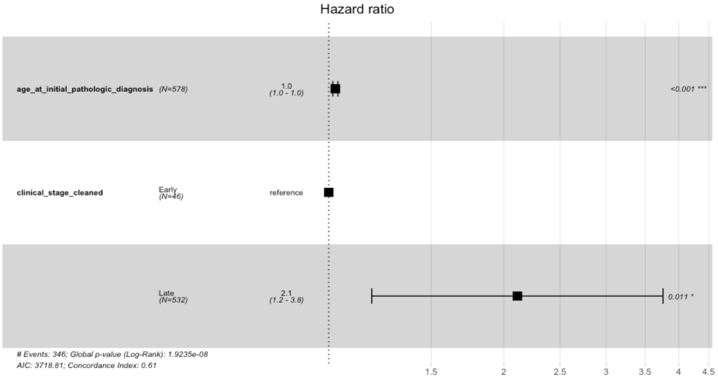
Figures A2.3-A2.6. Multivariate overall survival of the base model using covariates age at diagnosis, gender, and tumor stage for the BRCA, OV, COAD, and KIRC datasets. The age at dx and tumor stage were significant for all tumor types at p<0.05 level, where older age and late stage are associated with worse outcomes.

BRCA dataset: age at dx + gender + stage



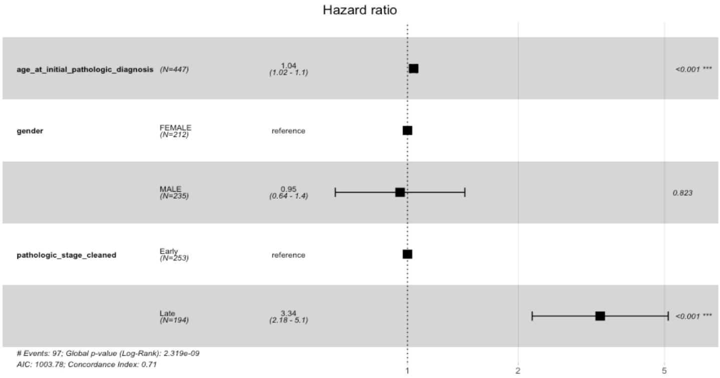
Events: 130; Global p-value (Log-Rank): 9.7611e-13
AIC: 1422.38; Concordance Index: 0.74

OV dataset: age at dx + gender + stage



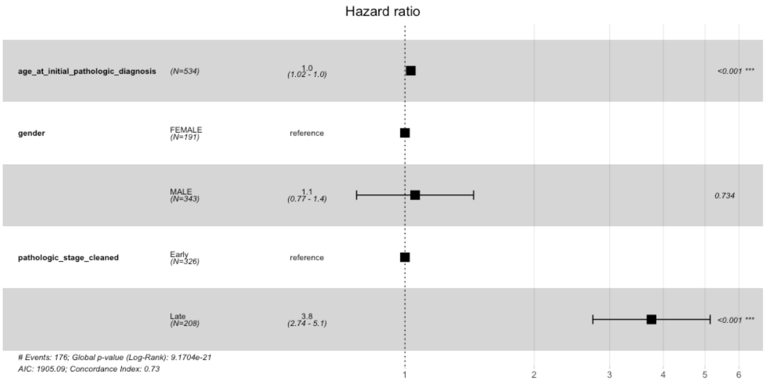
Events: 346; Global p-value (Log-Rank): 1.9235e-08
AIC: 3718.81; Concordance Index: 0.61

COAD dataset: age at dx + gender + stage



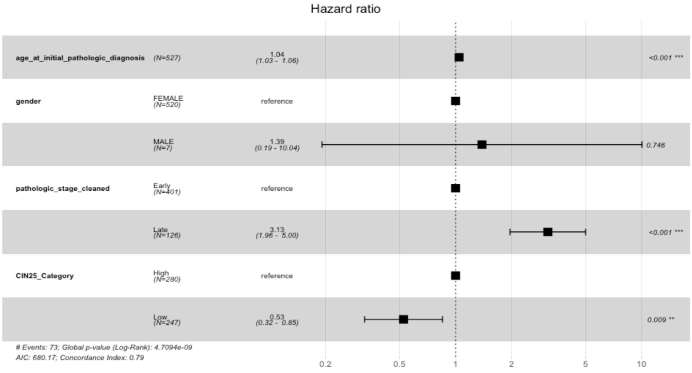
Events: 97; Global p-value (Log-Rank): 2.319e-09
AIC: 1003.78; Concordance Index: 0.71

KIRC dataset: age at dx + gender + stage

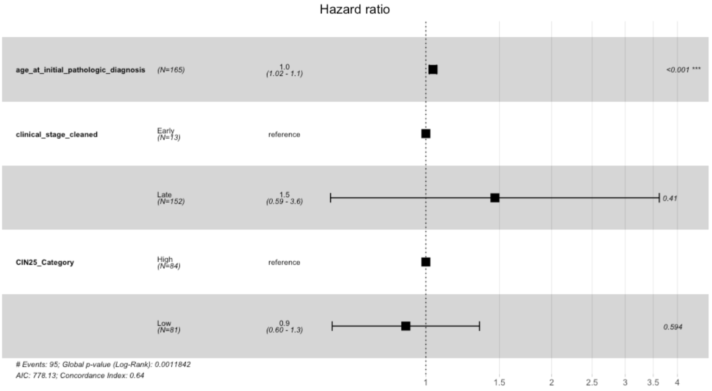


Figures A2.7-A2.10. Multivariate overall survival using covariates age at diagnosis, gender, tumor stage, and CIN25 category for the BRCA, OV, COAD, and KIRC datasets. The CIN25 category was significant for the BRCA and KIRC datasets at p<0.05 level, where lower score was associated with better outcomes (HR=0.53, p=0.009; 0.44, p=0.002). The age at dx was significant for all tumor types. The tumor stage was significant for BRCA, COAD, and KIRC datasets.

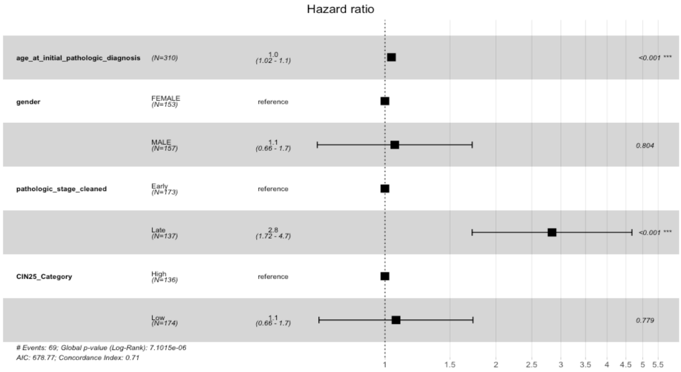
BRCA dataset: age at dx + gender + stage + CIN25 category



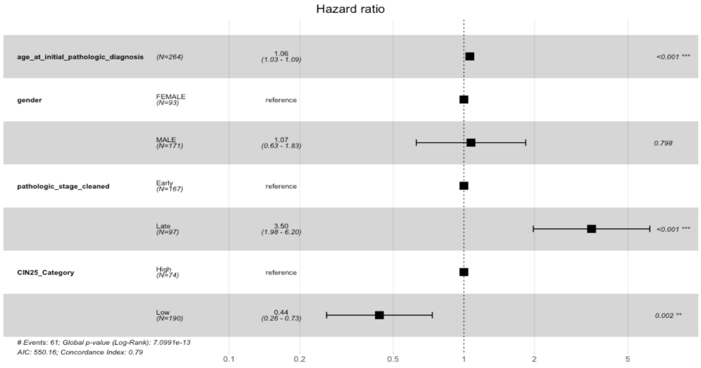
OV dataset: age at dx + gender + stage + CIN25 category



COAD dataset: age at dx + gender + stage + CIN25 category

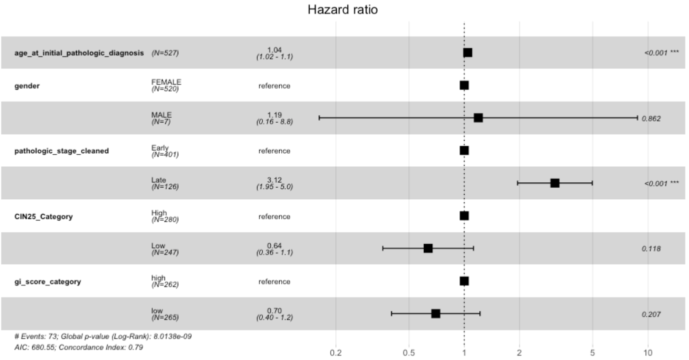


KIRC dataset: age at dx + gender + stage + CIN25 category

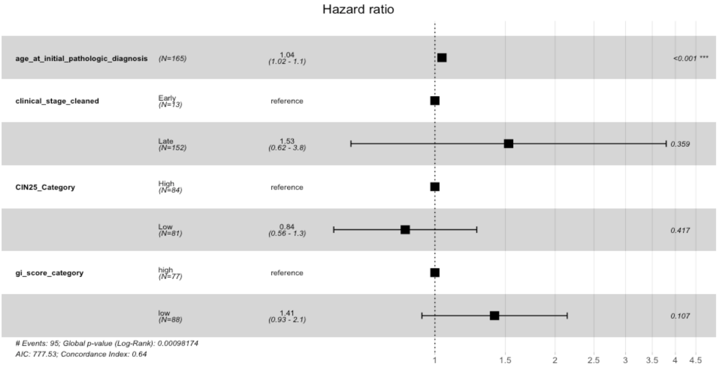


Figures A2.11-A2.14. Multivariate overall survival using covariates age at diagnosis, gender, tumor stage, CIN25 category, and GI category for the BRCA, OV, COAD, and KIRC datasets. The CIN25 category was significant for the KIRC dataset at $p<0.05$ level, where lower score was associated with better outcomes (0.44, p=0.002). The age at dx was significant for all tumor types. The tumor stage was significant for BRCA, COAD, and KIRC datasets.

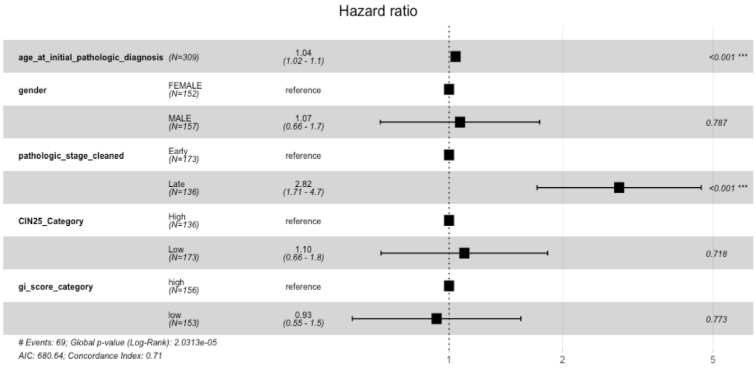
BRCA data: age at dx + gender + stage + CIN25 category + GI category



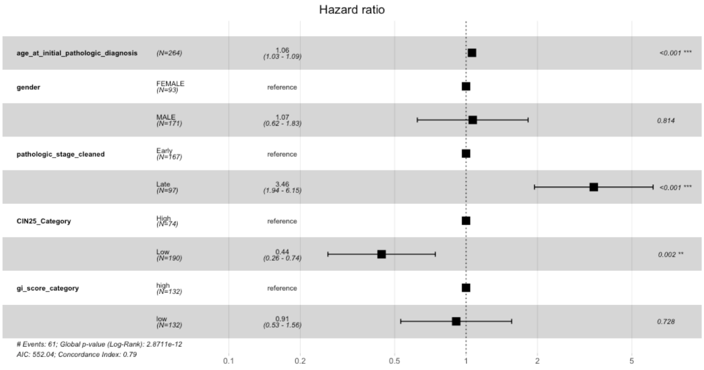
OV data: age at dx + gender + stage + CIN25 category + GI category



COAD data: age at dx + gender + stage + CIN25 category + GI category

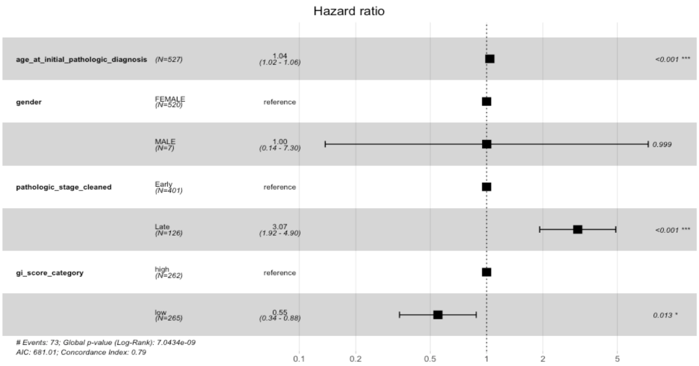


KIRC data: age at dx + gender + stage + CIN25 category + GI category

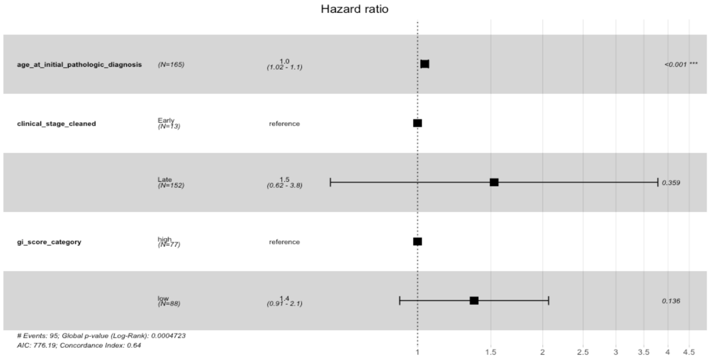


Figures A2.15-A2.18. Multivariate overall survival using covariates age at diagnosis, gender, tumor stage, and GI category for the BRCA, OV, COAD, and KIRC datasets. The GI category was significant for the BRCA dataset at p<0.05 level, where lower score was associated with better outcomes (0.55, p=0.013). The age at dx was significant for all tumor types. The tumor stage was significant for BRCA, COAD, and KIRC datasets.

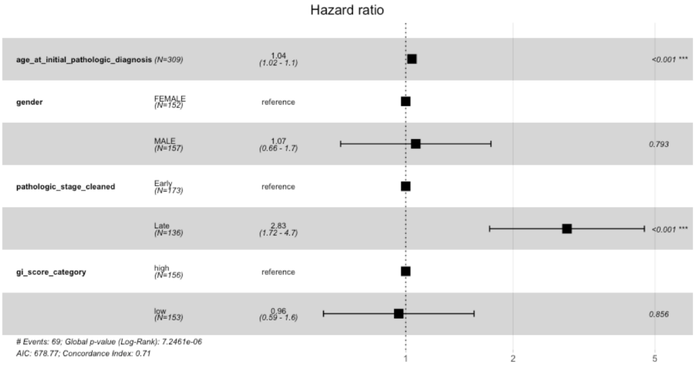
BRCA data: age at dx + gender + stage + GI category



OV data: age at dx + gender + stage + GI category



COAD data: age at dx + gender + stage + GI category



KIRC data: age at dx + gender + stage + GI category

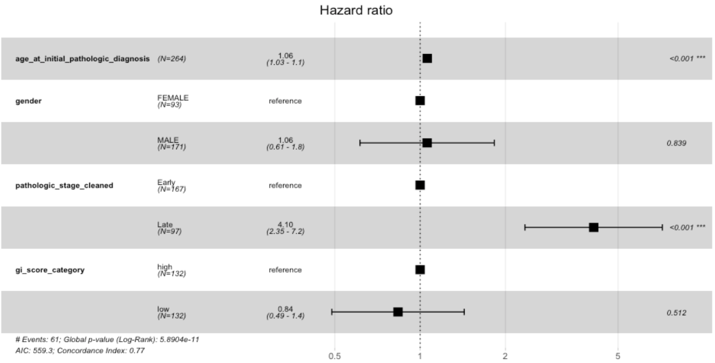
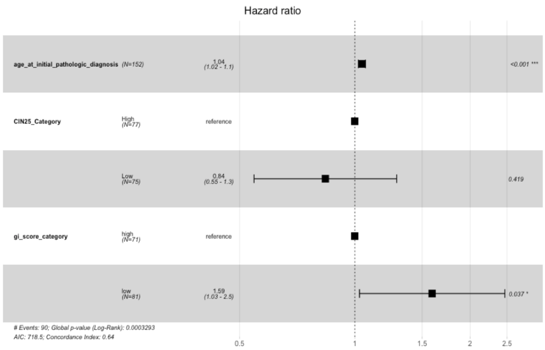


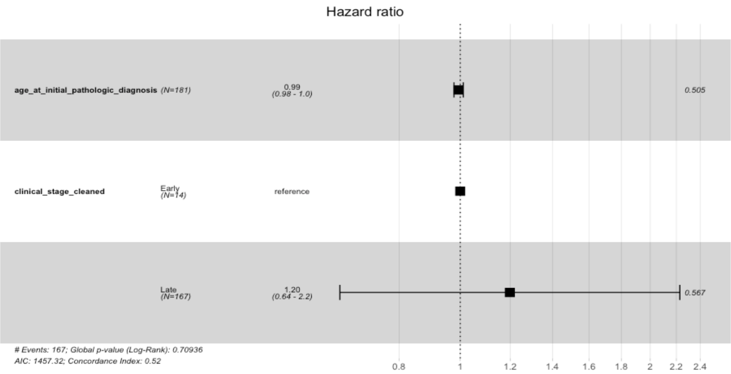
Figure A2.19. Multivariate analysis for late stage ovarian tumors using covariates age at diagnosis, gender, tumor stage, CIN25 category, and GI category. The GI score was significant after adjusted for age (HR=1.59, p=0.037).

Ovarian Late Stage: age at dx + CIN25 category + GI category

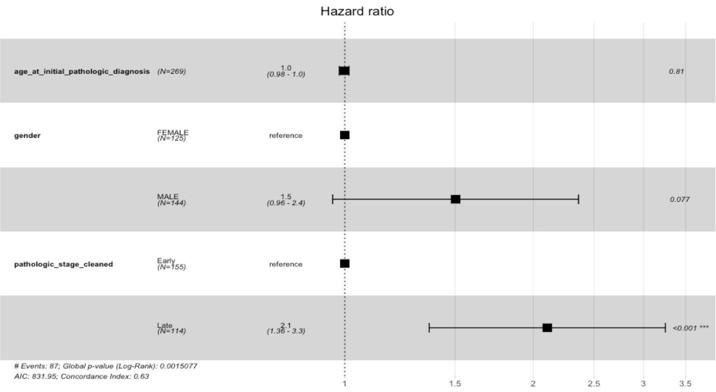


Figures A2.20-A2.22. Multivariate progression-free survival for the base model using covariates age at diagnosis, gender, and tumor stage for the BRCA, OV, COAD, and KIRC datasets. The tumor stage was significant for all tumor types, at p<0.05, where later stage was associated with worse outcomes. Being male was associated with worse outcomes in the KIRC dataset (HR=3.9, p=0.014).

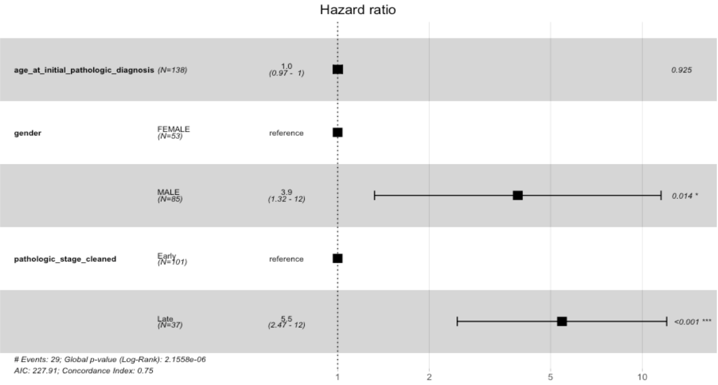
OV dataset: age at dx + gender + stage



COAD dataset: age at dx + gender + stage

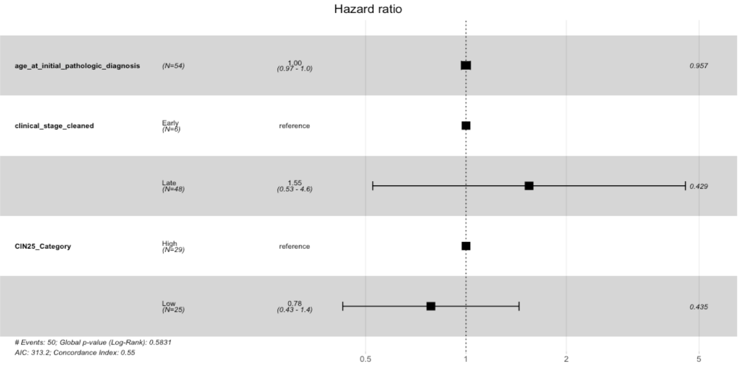


KIRC dataset: age at dx + gender + stage

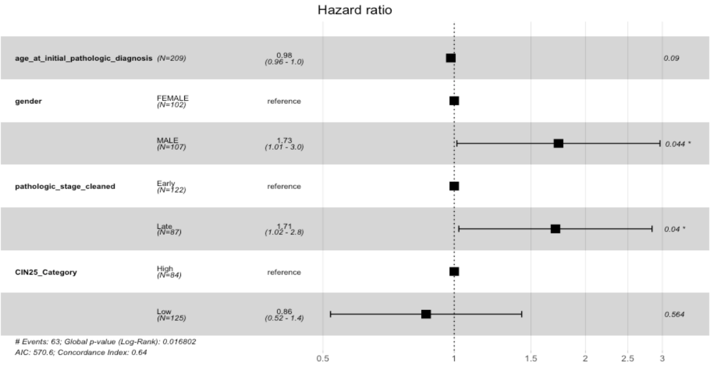


Figures A2.23-A2.25. Multivariate progression-free survival for the base model using covariates age at diagnosis, gender, tumor stage, and CIN25 category for the BRCA, OV, COAD, and KIRC datasets. The tumor stage was significant for all tumor types, at p<0.05, where later stage was associated with worse outcomes. Being male was associated with worse outcomes in the COAD and KIRC datasets (HR=0.55, p=0.045; HR=16, p=0.008).

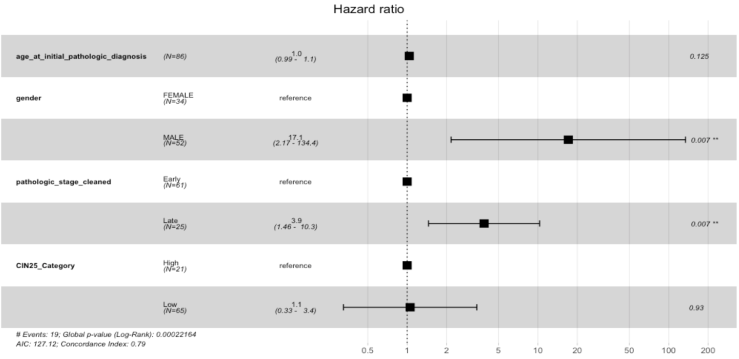
OV dataset: age at dx + gender + stage +
CIN25 category



COAD dataset: age at dx + gender + stage +
CIN25 category



KIRC dataset: age at dx + gender + stage + CIN25 category


Hazard ratio
age_at_initial_pathologic_diagnosis
(N=86)
(0.99 - 1.1)
gender
FEMALE
(N=34)
reference
MALE
(N=52)
(2.17 - 134.4)
0.007 **
pathologic_stage_cleaned
Early
(N=61)
reference
Late
(N=25)
(1.46 - 10.3)
0.007 **
CIN25_Category
High
(N=21)
reference
Low
(N=65)
(0.33 - 3.4)
# Events: 19; Global p-value (Log-Rank): 0.00022164
AIC: 127.12; Concordance Index: 0.79