Lizelle Winkelströter Correia

INFLUENCE OF SOMATIC MUTANT ALLELIC IMBALANCE IN BREAST CANCER CLINICAL CHARACTERISTICS

Departamento de Ciências Biomédicas e Medicina

2019

Lizelle Winkelströter Correia

INFLUENCE OF SOMATIC MUTANT ALLELIC IMBALANCE IN BREAST CANCER CLINICAL CHARACTERISTICS

Master in Oncobiology – Molecular Mechanisms of Cancer

Work under the supervision of:

Ana Teresa Maia, Ph.D

Sofia Braga, Ph.D

UAlg
UNIVERSIDADE DO ALGARVE

Departamento de Ciências Biomédicas e Medicina

2019

INFLUENCE OF SOMATIC MUTANT ALLELIC IMBALANCE IN BREAST CANCER CLINICAL CHARACTERISTICS

Authorship Statement

I hereby declare to be the author of this work, which is original and unpublished. Authors and papers consulted are duly cited in the text and are listed in the included references.

_____

Lizelle Winkelströter Correia

Live as if you were to die tomorrow.

Learn as if you were to live forever.

(Mahatma Gandhi)

Acknowledgments

Gostaria de agradecer à minha orientadora Professora Doutora Ana Teresa Maia por ter me aceite em seu grupo, por ter acreditado em mim e me apoiado por todo o percurso. Agradeço também à minha coorientadora Professora Doutora Sofia Braga e aos colegas de grupo Ramiro Magno, Joana Xavier, Isabel Duarte, Catarina Martins, Ana Fernandes, Juliana Machado e André Duarte pelo apoio e pelas ideias.

Meu muito obrigada à amiga Filipa Esteves pelos momentos de descontração, de risos, de desabafos, de ajuda e de orientação. O caminho teria sido muito mais difícil sem ti.

Às amigas Andréia, Vanessa, Adriana e Rafaela, muito obrigada por estarem sempre presentes. Guardo todos os momentos no coração. Obrigada aos amigos com quem dividi casa, Cândida e Daniel, pela companhia e risadas.

Agradeço, também aos amigos e familiares que deixei no Brasil e que mesmo de longe participam da minha vida e me apoiam. Não seria possível percorrer esse longo caminho sem o apoio e compreensão de vocês.

Agradeço aos meus avós pelo amor, dedicação e incentivo.

Obrigada ao meu irmão por todos os momentos ao longo das nossas vidas, você é um dos meus pilares de sustentação. Agradeço, também, à minha cunhada pelo apoio e carinho.

Obrigada à Sissi por ser a alegria constante que enche a casa de felicidade.

Meu agradecimento especial aos meus pais pelo amor e apoio incondicionais. Obrigada por investirem e acreditarem em mim e por sempre me incentivarem a ser mais e melhor. Todas as minhas vitórias só são possíveis por causa de vocês e são tão suas quanto minhas.

Agradeço à Deus todos os dias por ter posto pessoas tão maravilhosas em minha vida.

Abstract

Breast cancer is genomic and clinically heterogeneous, which represents a major challenge for treatment choice and has important consequences in survival rates. In the past decade studies profiled breast tumours, cataloguing somatic mutations and copy number alterations, which greatly improved prognosis and both expanded and facilitated treatment choices. In spite of this substantial progress, the current classification systems and biomarkers still do not encompass all breast cancer heterogeneity.

Dosage changes of mutated allele has been associated with alterations in prognosis and in response to treatment. However, all studies examined the clinical association of somatic mutations and allelic imbalance solely at DNA level, leaving the contribution of gene expression regulation somewhat overlooked. Cis-regulatory variation is known to be a major determinant of allelic imbalance of expression, and a vast majority of the human genome is known to be affected by this. The most robust approach to detect the effect of, and map, cis-regulatory variation is to analyse differential allelic expression in heterozygotes. So, I hypothesised that differential allelic expression of mutated genes contributes to breast cancer biology and outcome, by creating imbalances in the allelic levels of gene expression (dosage) of somatic mutations.

Using data from two independent sets of breast tumours, from METABRIC and TCGA projects, I studied the allelic imbalance of the two most mutated genes, *PIK3CA* and *TP53*. I show that preferential expression of the mutated allele presents positive selection, even more significantly that copy-number alterations. Also, I show that somatic mutations' allelic imbalance resulting from cis-regulation impacts on the clinical characteristics of tumours, namely by showing association with known prognostic factors. Finally, tumours with differential allelic expression of *PIK3CA's* mutated allele associated with poorer prognosis.

The work presented in this thesis suggests that allelic imbalance generated by cis-regulation has an essential role in modulating of the effect of somatic mutations in tumours, as well as tumour clinical behaviour.

Keywords: *PIK3CA*, *TP53*, Differential Allelic Expression, Breast Cancer, Cis-regulation, Allelic Imbalance

Resumo

O cancro da mama é uma doença heterogênea do ponto de vista clínico e genómico, o que representa um grande desafio na escolha do tipo de tratamento, com importantes consequências na taxa de sobrevida dos pacientes. Nas últimas décadas, estudos traçaram o perfil dos tumores da mama, catalogando mutações somáticas e alterações de número de cópias dos genes, melhorando muito o prognóstico e expandindo e facilitando as escolhas de tratamento dos doentes. Apesar desse progresso substancial, os sistemas de classificações e os biomarcadores atuais ainda não conseguem englobar toda a heterogeneidade do cancro da mama.

Dosagem variável do alelo mutado em relação ao alelo referência de oncogenes em cancros, variando desde pequenos a grandes desequilíbrios alélicos, tem sido associada a alterações em prognóstico e tratamento. No entanto, todos os estudos que abordaram o desequilíbrio alélico, fizeram-no apenas ao nível do DNA, negligenciando o papel da regulação da expressão dos genes. Sabe-se que variações na cis-regulação são grandes determinantes de expressão diferencial entre alelos e que a maior parte do genoma humano é afetado pela cis-regulação. Já foi também demonstrado que a cis-regulação afeta a penetrância das mutações, principalmente em contexto de portadores de mutações germinativas. No caso de mutações germinativas dos genes *BRCA1* e *BRCA2* em paciente com cancro da mama, foi descoberto que a maior expressão do alelo não mutado, capaz de produzir proteína para realizar as funções normais de reparação dos danos no DNA, associava a uma baixa penetrância, ou seja, era protector em relação ao desenvolvimento de cancro da mama.

Contudo, a forma como a variação cis-regulatória impacta a biologia tumoral e o desfecho clínico ainda segue inexplorado. Dessa forma, criou-se a hipótese que a expressão diferencial de alelos de genes com mutações somáticas contribui para biologia do cancro da mama e tem impacto na sobrevida dos pacientes. Nomeadamente, mutações somáticas em oncogenes e genes supressores de tumores cuja expressão que seja modelada por variação cis-regulatória, terão diferentes impactos na biologia e comportamento clínico do tumor que os comporta.

Para testar a hipótese, foram utilizados dados de sequenciação de nova geração de DNA e RNA de cancros de mama, e dados clínicos dos pacientes de dois grandes projectos, METABRIC e TCGA. Como os dois genes mais mutados e que apresentavam expressão

diferencial entre o alelo mutado e o alelo referência, nos dois projectos, foram os genes *PIK3CA* e *TP53*, optou-se por prosseguir o estudo analisando estes dois genes. Para a realização das análises foram calculados três rácios entre o alelo mutado e o alelo referência: rácio α que utiliza a contagem dos alelos no RNA e reflete a expressão total dos alelos; rácio β que utiliza a contagem dos alelos no DNA e reflete as alterações em número de cópias dos genes a nível do DNA; e o rácio γ em que o rácio α é normalizado pelo rácio β, refletindo assim a parte da expressão cujo responsável é a cis-regulação.

Ao analisar tumores com mutações somáticas no gene *PIK3CA*, foi observada associação de expressão diferencial da mutação com características do tumor, nomeadamente status dos recetores de estrogénio e progesterona e do HER2, com expressão preferencial do alelo mutado em tumores recetores de estrogénio e de progesterona negativos e HER2 positivo. Esta associação foi observada tanto a nível de expressão global do gene quanto quando se considerou apenas os efeitos da cis-regulação.

A análise de sobrevidas global e específica por cancro da mama demonstrou uma sobrevivência mais curta das pacientes com expressão diferencial entre alelo mutado e alelo referência do gene *PIK3CA*, quando comparado com pacientes que expressão aproximadamente as mesmas quantidades dos dois alelos. Ao avaliar as pacientes subdivididas de acordo com status de recetores hormonais, identificou-se um possível subgrupo de pior prognóstico dentro do grupo de tumores recetores de estrogénio e de progesterona positivos e HER2 negativo, que são tumores de melhor prognóstico.

Como, por pesquisa anterior do meu grupo, havia evidências de que o polimorfismo de nucleótido único (SNP) rs2699887 era um possível responsável pela cis-regulação do gene *PIK3CA* em tecido de mama normal, foi analisada a associação deste com desfecho clinico de tumores de mama e a única associação encontrada foi uma pior sobrevida do grupo de pacientes cujo tumor expressa o alelo menos frequente (alelo T) no subgrupo de tumores recetor de progesterona negativo.

Foi também observada associação entre expressão diferencial de alelo mutado e alelo referência de gene *TP53* com características biológicas do tumor, porém esta associação só foi vista nas amostras oriundas do projeto TCGA e apenas com status de recetores de estrogénio e de progesterona, nesses dois casos com expressão preferencial do alelo mutado em tumores recetores de estrogénio e de progesterona negativos. A análise se sobrevida global demonstrou associação de expressão diferencial do alelo *TP53* mutado com pior sobrevida quando todas as

amostras foram analisadas em conjunto e também nos subgrupos de tumores recetor de estrogénio positivo e de progesterona negativo.

Nas análises dos dois genes, *PIK3CA* e *TP53*, tanto para TCGA como METABRIC, quando comparada a dosagem de cada alelo seja por cis-regulação da expressão génica (RNA) ou alteração do número de cópias (DNA), observou-se que a maioria dos tumores expressa preferencialmente o alelo mutado, apesar de a nível de DNA possuir maior número de alelo referência, sugerindo uma seleção positiva da expressão do alelo mutado em tumores de mama. Se bem que a análise do *PIK3CA*, um oncogene típico, foi restringida a mutações *missense*, dado que o gene *TP53* pode actuar como gene supressor de tumores ou oncogene consoante a mutação adquirida, a sua análise foi ainda extendida ao tipo de mutação presente em cada tumor. O que verificámos foi que a selecção positiva significava ainda uma separação clara entre mutações *missense*, associadas ao seu papel de oncogene e que verificámos estarem mais associadas a expressão preferencial do alelo mutado, e mutações *frame-shift*, normalmente associadas ao seu papel de gene supressor de tumores e que verificámos estar associado a expressão preferencial do alelo referência.

O trabalho realizado no âmbito da tese aqui apresentada providencia, pela primeira vez, evidências que apoiam o papel essencial do desequilíbrio alélico gerado pela cis-regulação, além do gerado pela alteração no número de cópias, na modulação do efeito das mutações somáticas nos tumores. Estes achados suportam a realização de pesquisa de níveis de expressão de alelos de mutações somáticas oncogénicas como parte do manejo clínico de pacientes com cancro da mama. Para além disso, os achados também suportam novos estudos sobre a forma como estes desequilíbrios da expressão génica podem estar ativando ou silenciando outros genes relacionados com o cancro.

Palavras chave: *PIK3CA*, *TP53*, expressão diferencial de alelos, cancro da mama, cis-regulação, desequilíbrio alélico

Table of Contents

List of Figures

List of Tables

Abbreviations

AKT	Protein kinase B
ATR	Ataxia Telangiectasia and Rad3-Related Protein
CBP	CREB-binding protein
CCND1	Cyclin D1
CCND3	Cyclin D3
CHEK2	Gene that codes checkpoint kinase 2 (CHK2)
CI	Confidence Interval
CREB	cAMP response element binding protein
DAE	Differential Allelic Expression
DNA	Deoxyribonucleic acid
DNAseq	DNA sequencing
DSS	Disease Specific Survival
EGFR	Epidermal growth factor receptor
EIF4EBP1	Eukaryotic Translation Initiation Factor 4E Binding Protein 1
EMSY	BRCA2-interacting transcriptional repressor
ER	Oestrogen receptor
ERBB2	Gene that codes protein HER2
FRAP	Gene that codes protein FKBP12-rapamycin complex-associated protein
GATA3	GATA binding protein 3
GOF	gain-of-function
GPX4	Glutathione peroxidase 4
GWAS	genome-wide association study
HER1	human epidermal growth factor receptor 1, also known as EGFR
HER2	human epidermal growth factor receptor 2
HER3	human epidermal growth factor receptor 3
HER4	human epidermal growth factor receptor 4
HoxA5	Homeobox protein Hox-A5

HR	Hazard Ratio
IntClust	Integrative Cluster
LD	Linkage Disequilibrium
MADE group	Mutant Allele Differential Expression; $\|\gamma\| \geq 0.58$
MADE_mut group	$\gamma \geq 0.58$, or preferential expression of mutant allele
MADE_wt group	$\gamma \leq -0.58$, or preferential expression of wild-type allele
MAP2K4	mitogen-activated protein kinase kinase 4
MAP3K1	Mitogen-Activated Protein Kinase Kinase Kinase 1
MAPK	mitogen-activated protein kinases
MAPK1	Mitogen-Activated Protein Kinase 1
MASI	mutant allele specific imbalance
MDM2	Mouse double minute 2 homolog
METABRIC	The Molecular Taxonomy of Breast Cancer International Consortium
miRNA	micro-RNA
mRNA	Messenger RNA
MTHFR	methylenetetrahydrofolate reductase
mTOR	mammalian target of rapamycin
mut=wt group	Normalized expressions of mutated and wild-type alleles were not different
NGS	Nottingham Grading System
OS	Overall Survival
PAM50	Prediction Analysis of Microarray 50
PDK1	3-phosphoinositide-dependent protein kinase-1
PH	Proportional hazard
PI3K	Phosphoinositide 3-kinases
PIK3CA	Phosphatidylinositol 4,5-bisphosphate 3-kinase catalytic subunit alpha isoform
PIK3R1/2/3	Phosphoinositide-3-Kinase Regulatory Subunit 1/2/3
PIP	phosphatidylinositol 3-phosphate
PIP2	phosphatidylinositol 3,4-bisphosphate
PIP3	phosphatidylinositol 3,4,5-trisphosphate

PR	Progesterone receptor
PTEN	Phosphatase and Tensin Homolog
RB1	Retinoblastoma 1
RNA	Ribonucleic acid
RNAseq	RNA sequencing
RPS6KB1/2	ribosomal protein S6 kinase B1/2
rSNP	Regulatory single nucleotide polymorphism
RTK	Receptor tyrosine kinase
RUNX1	runt-related transcription factor 1
SNP	Single nucleotide polymorphism
TAD	transactivation domain
TCGA	The Cancer Genome Atlas
TNBC	triple (ER/PR/HER2) negative breast cancer
TNM system	tumour, node, metastasis staging system
TRXR2	thioredoxin reductase-2
TSC1/2	Tuberous Sclerosis Complex 1/2
α	α = Log2[(mutant allele count in RNA)/(wild-type allele count in RNA)]
α_DAE group	$\lvert\alpha\rvert \geq 0.58$
α_DAEmut group	$\alpha \geq 0.58$, or preferential expression of mutant allele
α_DAEwt group	$\alpha \leq -0.58$, or preferential expression of wild-type allele
α_noDAE group	Overall expressions of mutated and wild-type alleles were not different
β	β = Log2[(mutant allele count in DNA)/(wild-type allele count in DNA)]
γ	the normalized mutant allele expression ratio $\gamma = [\alpha]–[\beta]$, which is a measure of mutant allelic expression imbalance due to cis-regulation alone

CHAPTER 1
INTRODUCTION

1 Introduction

1.1. Cancer

Cancer is a disease of malfunctioning cells, in which the architecture is less organized than that of nearby normal cells, and the regulatory circuits that govern normal cell proliferation and homeostasis are defective (Hanahan and Weinberg, 2000; Weinberg, 2013). Tumorigenesis in humans is a multistep process reflecting the genetic alterations that drive the progressive transformation of normal human cells into highly malignant derivatives (Hanahan and Weinberg, 2000). This process could be rationalized by the need of incipient cancer cells to acquire the traits that enable them to become tumorigenic and ultimately malignant.

In 2011, Hanahan and Weinberg updated their model on tumorigenesis, with a total of ten biological capabilities acquired during the multistep development of human tumours (Figure 1.1).The original six capabilities that most if not all cancers have acquired during their development, albeit through various mechanistic strategies, include sustaining proliferative signalling, evading growth suppressors, resisting cell death, enabling replicative immortality, inducing angiogenesis, and activating invasion and metastasis. To those, four new capabilities were added: genome instability, inflammation, reprogramming of energy metabolism and evading immune destruction. These hallmarks allow cancer research to deal with cancer phenotypic complexities as manifestations of a small set of underlying organizing principles, increasingly affecting the development of new means to treat human cancer (Hanahan and Weinberg, 2011).

Cancer incidence and mortality are rapidly growing world-wide. The reasons are complex but reflect both aging and growth of the population, as well as changes in the prevalence and distribution of its main risk factors, several of which are associated with socioeconomic development. For many cancers, incidence rates are generally 2-fold to 3-fold higher in transitioned compared with transitioning economies. However, the differences in mortality between these two regions are smaller, in part because of a higher case fatality for many cancer types in transitioning countries. Although, Europe accounts for 23.4% of the total cancer cases and 20.3% of the cancer deaths, it represents only 9% of the global population (Bray *et al.*, 2018).

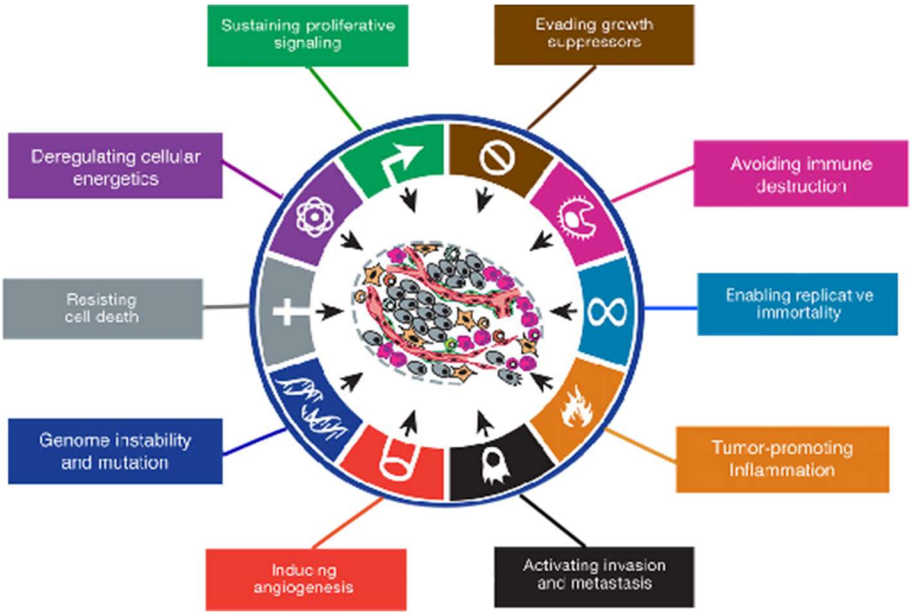


Figure 1.1 The Hallmarks of Cancer. This illustration encompasses the ten hallmark capabilities of cancer (Adapted from Hanahan and Weinberg, 2011).

1.1.1 Breast Cancer

1.1.1.1 Epidemiology

Breast cancer is a complex and heterogeneous multifactorial disease with both environmental and genetic risk factors (Véron, Blein and Cox, 2014).

In women, incidence rate for breast cancer far exceeds that of other cancers in both transitioned and transitioning countries, followed by colorectal cancer in transitioned countries, and cervical cancer in transitioning countries (Figure 1.2). Worldwide, there were about 2.1 million newly diagnosed female breast cancer cases in 2018, accounting for almost 1 in 4 cancers among women. Incidence rates of breast cancer have been rising for most countries over the last decades, with some of the most rapid increases occurring where rates have been historically relatively low, as for example, in transitioning countries in South America, Africa, and Asia. Along with lung and colorectal cancers, female breast cancer explains one-third of the cancer incidence and mortality burden worldwide, and is the third cancer in terms of incidence and the fifth in terms of mortality, when accounting for males and females (Bray *et al.*, 2018). Even though breast cancer can occur in men, it is more than 100 times more common in women (Makki, 2015).

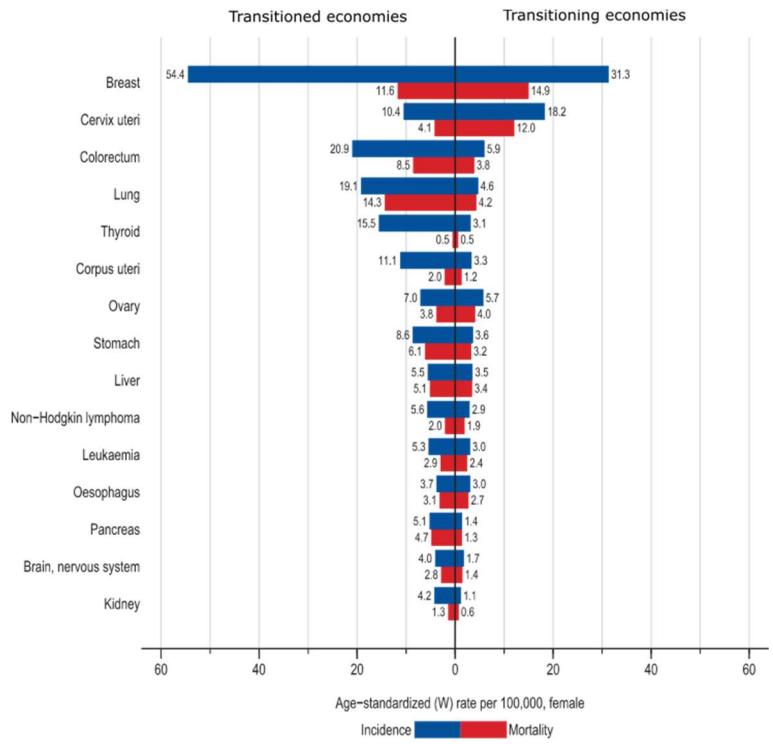


Figure 1.2: Bar Charts of Incidence and Mortality Age-Standardized Rates in Transitioned Economies Regions Versus Transitioning Economies Regions Among Women in 2018. The 15 most common cancers world (W) in 2018 are shown in descending order of the overall age-standardized (Adapted from Bray *et al*., 2018).

In Portugal breast cancer is the most commonly diagnosed cancer in women, with an estimated 6088 new cases and 1570 deaths in 2012, accounting for 30% of all cancer cases and 16% of all cancer deaths (Bray *et al.*, 2018; Forjaz de Lacerda *et al.*, 2018). It is estimated that one in each 11 women in Portugal will be diagnosed with breast cancer throughout their lives (Geroge, 2012).

1.1.1.2 Aetiology

Breast cancer is a malignant proliferation of epithelial cells lining the ducts or lobules of the breast. (Kasper *et al.*, 2015). Breast cancer starts when cells in the breast begin to grow out of control, and the tumour becomes malignant if the cells can invade surrounding tissues or spread (metastasize) to distant areas of the body (Bland and Copeland, 2009).

Breast cancer is a hormone-dependent disease with increased incidence associated with age, but the increase slows down after the age of menopause (median age of menopause 52 years). Women with latter menarche, earlier first pregnancy or menopause before 50 years of

age have lower breast cancer risk. Also the duration of maternal nursing correlates with substantial risk reduction with longer duration associated with smaller risk (Kasper *et al.*, 2015).

Exogenous hormones also have a role in breast cancer aetiology, with studies suggesting hormonal contraceptives cause a small increase in breast cancer risk and hormonal replacement therapy being associated with an important increase in its incidence. In the past decade, the decrease in the number of women on hormonal replacement therapy has already led to a consistent decrease in breast cancer incidence (Kasper *et al.*, 2015), which supports this association.

Hereditary disease accounts only for 5-10% of all breast tumours, whereas the majority of breast cancers are sporadic, resulting from the accumulation of acquired somatic alterations, that are individually rare and encountered in unique combinations in every cancer (Santarpia *et al.*, 2016; Valencia *et al.*, 2017).

Currently identified breast cancer susceptibility genes and alleles can be stratified by their conferred risk in high, moderate and low-penetrant categories. *BRCA1* and *BRCA2* are the two most commonly mutated high-penetrance genes, and about 20– 30% of the familial breast cancer risk is attributable to germline mutations in one of these two genes. Although germline mutations in *PTEN, TP53, STK11*, and *CDH1* also confer a high breast cancer risk, they are very rare and mostly found within the context of the familial cancer syndromes they cause. Hence, mutations in these genes explain no more than 1% of the familial breast cancer risk. A more intermediate risk of developing breast cancer is conferred by germline mutations in the genes *CHEK2, ATM, PALB2*, and *NBS1*, which are more prevalent in the general population than the high-risk mutations previously described. Together they explain another 5% of the familial breast cancer risk. Lastly, low penetrant breast cancer susceptibility alleles have been identified through large-scale genome wide association studies (GWAS), which explain about 18% of the familial breast cancer risk. In this way, approximately 50% of familial breast cancer risk is still unexplained (Eccles *et al.*, 2013; Rivandi, Martens and Hollestelle, 2018).

1.1.1.3 Clinical and Pathological Classification

The mainstays of breast cancer characterization are still histologic subtype, hormone receptor and HER2 statuses, histological grade, and tumour stage, which provide a basic reflection of degree of tumour differentiation and growth rate. These characteristics are

routinely used to predict outcomes in non-treated breast cancer patients and to determine the appropriate therapy, (Schnitt, 2010; Reisenbichler *et al.*, 2017; Provenzano, Ulaner and Chin, 2018).

Breast carcinomas are the most common malignant lesions diagnosed in the breast, although different types of sarcomas and lymphomas can also be diagnosed. Breast carcinoma is usually classified primarily by its histological appearance, originating from the inner lining epithelium of the ducts or the lobules that supply the ducts with milk (Makki, 2015). Seventy to 80% of breast cancers fall into the ductal/no-special-type category (invasive ductal carcinoma), which, rather than representing a unique entity, shows marked heterogeneity with respect to tumour morphology, molecular biology, and prognosis (Provenzano, Ulaner and Chin, 2018). The remaining 20 to 30% of breast carcinomas are classified into special types based on the dominating growth pattern. The main special types are lobular (7-16%), tubular (1-10%), medullary and mucinous (0.5-2%) carcinomas (Figure 1.3) (Bland and Copeland, 2009; Russnes *et al.*, 2017).

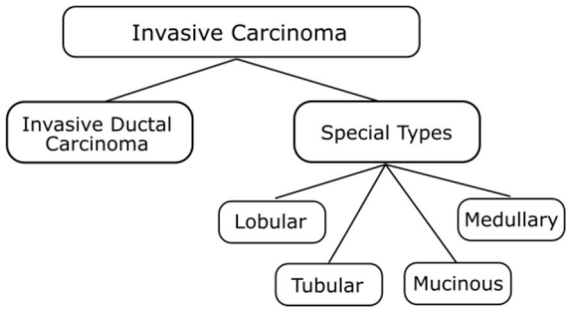


Figure 1.3*:* Histological classification of breast cancer subtypes. This scheme categorizes the heterogeneity found in breast cancer based on architectural features. (Adapted from Malhotra *et al*., 2010)

Histologic grading is more important than the morphologic type for clinical management of patients with breast cancer (Provenzano, Ulaner and Chin, 2018). The Nottingham (Elston-Ellis) modification of the Scarff -Bloom-Richardson grading system, also known as the Nottingham Grading System (NGS), is the grading system recommended by various international professional bodies (World Health Organization, American Joint Committee on Cancer, European Union, and the Royal College of Pathologists) to determine histological grade (NHS Cancer Screening Programmes and The Royal College of Pathologists, 2005; Rakha *et al.*, 2010; Amin *et al.*, 2017). In breast cancer, it refers to the semi-quantitative

evaluation of morphological characteristics and is a relatively simple and low-cost method. NGS is based on the degree of differentiation of the tumour tissue. It evaluates three morphological features: (a) degree of tubule or gland formation, (b) nuclear pleomorphism, and (c) mitotic count. Histological grade is an important determinant of breast cancer outcome accurately predicting tumour behaviour, particularly in earlier small tumours. Tumours of different histological grades show distinct molecular profiles at the genomic, transcriptomic, and immunohistochemical levels (Rakha *et al.*, 2010). It is also noted that breast tumours of high histological grade are generally large at presentation and are associated with local or distant metastasis, compared with tumours of low histological grade (Tao *et al.*, 2015).

Hormone receptor status (oestrogen and progesterone receptors) and HER2 status analyses are another important part of the initial diagnosis of breast cancer, because they have predictive and prognostic value. Up to 80% of breast cancers express oestrogen receptor (ER) and 55% to 65% are positive for progesterone receptor (PR) expression (Fragomeni, Sciallis and Jeruss, 2018). A major mechanism of action of ER is through its function as a transcription factor (Nagaraj and Ma, 2015). Nuclear ER activation by its ligand oestrogen initiates transcription and translation of proteins involved in cell division, angiogenesis, and survival of breast cells (Keegan *et al.*, 2018). The gene expression of PR is dependent on ER and consequently PR expression has been considered to indicate an intact ER response pathway (Yip and Rhodes, 2014). While ER can stimulate the expression of PR, PR in the presence of progesterone has been shown to interact with and alter the location at which ER binds to chromatin. The altered chromatin binding of ER results in a switch from regulating genes implicated in proliferation to modulating genes associated with cell cycle arrest, apoptosis and differentiation (Nicolini, Ferrari and Duffy, 2018). In ER positive patients, where the PR is absent (approximately 20% of ER positive breast cancers), overall survival, breast cancer-specific survival and disease-free survival are significantly poorer than for patients with tumours positive for both receptors (Yip and Rhodes, 2014). Patients whose tumours are ER positive benefit from endocrine therapy targeting ER, which has broadly focused on one of three approaches: blocking oestrogen bio-synthesis (e.g. anastrozole and other aromatase inhibitors), antagonizing ligand binding to ER (e.g. tamoxifen) or inducing ER downregulation (e.g. fulvestrant). Endocrine therapy can reduce local and distant recurrence and mortality (Fragomeni, Sciallis and Jeruss, 2018; Keegan *et al.*, 2018).

Approximately 10% to 15% of breast cancers will overexpress the human epidermal growth factor receptor 2 protein (HER2), a receptor tyrosine kinase that is involved in

regulation of cell growth, survival, adhesion, migration and differentiation (Moja *et al.*, 2012; Russnes *et al.*, 2017). About 10% of all ER positive breast tumours also reveal HER2 overexpression, while about half of breast cancers with HER2 overexpression co-express ER (Keegan *et al.*, 2018). The human epidermal growth factor family of tyrosine kinase receptors includes human epidermal growth factor receptor 1 (HER1; also known as epidermal growth factor receptor [EGFR]), HER2, HER3, and HER4. Most of these receptors can be activated by ligand-dependent dimerization — the pairing of two of the same receptors (homodimerization) or two different receptors (heterodimerization) — within the plasma membrane. Although there are no known endogenous ligands for HER2, when HER2 is overexpressed, it can undergo ligand-independent activation via spontaneous homodimerization and autoactivation and may facilitate heterodimerization with HER1 and HER3, resulting in unregulated growth and cell survival in breast cancer. HER2 overexpression (HER2 positivity) in breast cancer has been associated with poorer clinical outcomes compared with HER2 negative disease (Eroglu, Tagawa and Somlo, 2014; Keegan *et al.*, 2018). HER2 is regarded as a prognostic and predictive factor. HER2 positive tumours were previously associated with a poorer prognosis, but the advent of a range of anti-HER2 targeted therapies, such as trastuzumab and lapatinib, given with cytotoxic chemotherapy, has significantly improved the overall and disease free survival for patients with this subtype of breast cancer (Moja *et al.*, 2012; Russnes *et al.*, 2017; Keegan *et al.*, 2018).

While ER positive tumours have a better prognosis, triple (ER/PR/HER2) negative breast cancer (TNBC), which accounts for approximately 15% of all breast cancers, are associated with the poorest outcome (Di Leo *et al.*, 2015). There is no adjuvant therapy other than chemotherapy for TNBC because no target marker has been validated (Youn *et al.*, 2018).

One of the first priorities in patient management after histopathologic diagnosis of breast cancer is to determine the extent of disease, in a process called staging. Information obtained from staging is used to define the extent of the disease as localized, as exhibiting spread outside of the organ of origin to regional but not distant sites, or as metastatic to distant sites. The most widely used system of staging is the TNM (tumour, node, metastasis) system codified by the International Union Against Cancer and the American Joint Committee on Cancer. The TNM classification is an anatomically based system that categorizes the tumour on the basis of the size of the primary lesion (T1-4, where a higher number indicates a tumour of larger size), the presence of nodal involvement (N0 for the absence and N1-3 for the presence of involved nodes) and the presence of metastatic disease (M0 and M1 for the absence and presence of

metastases, respectively). The various permutations of T, N and M scores are then broken into stages, usually designated by the roman numerals I through IV. Tumour burden increases and curability decreases with increasing stage (Kasper *et al.*, 2015).

Although the current well-established clinical and histological factors show strong association with prognosis and outcome, there are increasing concerns that these variables are limited in their ability to capture the diversity of clinical behaviour of breast cancer and that they would not be sufficient to tailor the therapy to individual patients (Rakha *et al.*, 2010). One example is the observation that even though more than 75% of patients have ER positive invasive ductal carcinoma breast cancers, their outcomes and responses to therapy are extremely varied (Pereira *et al.*, 2016).

1.1.1.4 Molecular Classification

In order to refine breast cancer classification and to better assess prognosis and therapy choice, molecular techniques were used to study gene expression profiling, generating the molecular classification, which classifies tumours into intrinsic subtypes (Schnitt, 2010; Russnes *et al.*, 2017).

The seminal articles that led to the identification of five intrinsic subtypes emerged as a result of the work by Perou *et al.* and Sorlie *et al.* more than a decade ago, who analysed gene expression in breast carcinoma samples taken before and after chemotherapy (Perou *et al.*, 2000; Sorlie *et al.*, 2001; Russnes *et al.*, 2017).

The most reproducibly identified molecular intrinsic subtypes among the hormone receptor-positive cancers are the Luminal A and Luminal B groups (51-61% and 14-16% of breast cancers, respectively) and the HER2 and Basal-like groups are the major molecular subtypes identified among hormone receptor-negative breast cancers (7-9% and 11-20% of breast cancers, respectively) (Bland and Copeland, 2009; Russnes *et al.*, 2017). A Normal-like subtype that can be either hormone receptor positive or negative have also been identified in some studies, but, because it was developed by training on normal breast tissue, it is suspected that it is mainly an artefact of having a high percentage of normal cells "contamination" in the tumour specimen. Another possible explanation is a group of slow-growing Basal-like tumours that lack expression of the proliferation genes (Parker *et al.*, 2009).

In order to facilitate clinical use of this classification, the Prediction Analysis of Microarray 50 (PAM50) classifier (Prosigna®), which is a commercial test, has been developed. It is based on the expression profiling of 50 genes, that can separate tumours into the five intrinsic subtypes, and has added significant prognostic and predictive value to pathologic staging, histologic grade, and standard clinical molecular markers (Parker *et al.*, 2009).

These breast cancer molecular subtypes differ not only with regard to their patterns of gene expression, but also with clinical features, response to treatment, and prognosis (Schnitt, 2010). There is a highly significant difference in overall survival between the subtypes, with the Basal-like and HER2 positive subtypes associated with the shortest survival times and relapse-free survival (Figure 1.4) (Sorlie *et al.*, 2001). Luminal A is overall the most frequently occurring breast cancer expression subtype in the population and, while associated with the highest median overall survival, is also characterized by the most variability in survival. Moreover, it has been shown that the risk of late mortality in this subtype persists at least over 10 years after initial diagnosis (Ciriello *et al.*, 2013).

These differences in survival, along with studies that revealed further complexity and diversity between and within the known subtypes, indicate that further sub-classification of patients into different treatment groups should be considered (Prat and Perou, 2011; Ciriello *et al.*, 2013; Norum, Andersen and Sørlie, 2014).

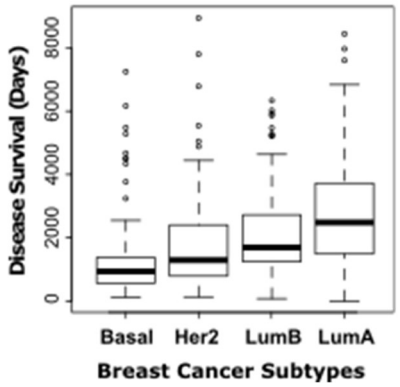


Figure 1.4*:* Boxplot statistics of disease survival for deceased patients from the METABRIC dataset across the four major PAM50 subtypes. (Adapted from Ciriello *et al*. 2013)

1.1.1.5 Mutations

Cancer is a genetic disease driven by DNA alterations, including chromosomal rearrangements, single-base mutations, and epigenetic changes resulting in activation of growth-promoting genes (oncogenes) or suppression of growth-inhibiting genes (tumour suppressor genes). As such, in the last decades, many efforts have been concentrated to supplement the morphological and molecular classifications of breast carcinoma with other molecular parameters that can provide a clearer appreciation of the heterogeneity of breast cancer and for better predicting tumour behaviour and improve therapeutic strategies. The understanding of mutations and pathways that drive tumorigenesis is a great part of this effort (Makki, 2015; Provenzano, Ulaner and Chin, 2018).

Mutations alter the information content of genes. Alternative versions of the same gene are called alleles. In a diploid cell, such as human cells, each autosomal gene has two copies/alleles (one inherited from the mother and one from the father), occupying the corresponding position (locus) on homologous chromosomes (Alberts *et al.*, 2014). If a mutation occurs resulting in a mutant allele in a germ cell (sperm and egg), it can be transmitted to offspring, and is said to occur via germ line. Mutations affecting the genomes of cells everywhere else in the body (somatic mutations) may well affect the cells in which they occur, but will have no prospect of being transmitted to the offspring of an organism (Weinberg, 2013).

Mutations strike the genome mostly randomly and only rarely hit critical genes that, when mutated, confer advantageous phenotypes leading to the clonal expansions that drive multistep tumorigenesis. Consequently, a cell that happens to acquire an advantageous mutant allele will also carry numerous other mutations that have struck other genes throughout its genome; these other mutant genes have no influence on cancer cell phenotype and consequently are irrelevant to tumour progression. The mutations that are critical participants in tumour progression are called driver mutations, whereas mutations in other non-important genes are passenger mutations. Driver mutations can be identified because they affect genes that are the objects of recurrent mutations (Weinberg, 2013).

The two most common gain-of-function events targeting oncogenes in human cancers are somatic mutations (single-nucleotide changes, small in-frame insertions and deletions) and focal DNA copy-number amplifications. Individually, both events are common, well-studied, and, in a growing number of cases, therapeutically actionable. Nevertheless, with few

exceptions, most recurrently mutated oncogenes are not focally amplified and most recurrent, focally amplified oncogenes are not frequently mutated (Bielski *et al.*, 2018).

Cancers arise through successive waves of clonal expansion dependent on the sequential acquisition of driver mutations. A central parameter of cancer development is therefore the number of driver mutations required for conversion of a normal cell into a symptomatic cancer. Estimates based on cancer age–incidence curves have indicated that approximately five rate-limiting steps underlie the development of common adult solid tumours, such as breast cancer. Stephens *et al.* showed somatic driver point mutations and/or copy number changes in at least 40 cancer genes were implicated in the development of 100 breast cancers. Seven of the 40 genes (*TP53*, *PIK3CA*, *ERBB2*, *MYC*, *FGFR1/ZNF703*, *GATA3* and *CCND1*) were mutated in more than 10% of the tumours (Stephens *et al.*, 2012). In another study in which somatic mutations were analysed in the whole genome of 560 breast cancers, the 10 most frequently mutated genes in breast cancer were *TP53*, *PIK3CA*, *MYC*, *CCND1*, *PTEN*, *ERBB2*, *FGFR1/ZNF703* locus, *GATA3*, *RB1* and *MAP3K1*, and these accounted for 62% of drivers (Nik-zainal *et al.*, 2016).

In The Cancer Genome Atlas (TCGA) project, which is a cancer genomic program that molecularly characterized primary cancers, three genes were mutated in more than 10% of cases *TP53* (37%), *PIK3CA* (36%) and *GATA3* (11%) and along with *MAP3K1* (8%), *MLL3* (7%) *CDH1* (7%), *MAP2K4* (4%), *RUNX1* (4%), *PTEN* (3%) were the top mutated genes (Koboldt, Fulton and McLellan, 2012). Depending on the study the order of the top mutated genes in breast cancer slightly change, but since *TP53* and *PIK3CA* are always listed as the top mutated, their pathways will be detailed bellow.

PI3K/AKT pathway

The PI3K/AKT signalling pathway is the most recurrently altered pathway in breast cancer, apparently with different biologic impact on specific cancer subtypes (Cossu-Rocca *et al.*, 2015).

Phosphoinositide 3-kinases (PI3Ks) belong to a conserved family of lipid kinases involved in vital functions such as cell division control, differentiation, cytoskeleton re-organization, and intracellular trafficking (Wang *et al.*, 2017). Three PI3K classes have been defined on the basis of their primary structure, regulation and in vitro lipid substrate specificity. Class I PI3Ks generate phosphatidylinositol (PI) 3-phosphate (PIP), PI 3,4-bisphosphate (PIP2) and PI 3,4,5-trisphosphate (PIP3). Class II PI3Ks generate PIP and PIP2, while class III PI3Ks

produce PIP only (Leevers, Vanhaesebroeck and Waterfield, 1999; Lux *et al.*, 2016). Class I can be further divided into class IA and class IB, based on their activation receptors: receptor tyrosine kinase (RTKs) activate Class IA proteins and G-protein- coupled-receptors activate Class IB PI3Ks. Thus far, only the class IA PI3Ks have been implicated in human cancer (Yuan and Cantley, 2008; Wang *et al.*, 2017).

Class IA PI3Ks are heterodimers consisting of a catalytic (p110) and a regulatory (p85) subunit, with the latter stabilizing the former in quiescent cells and suppressing PI3K activity. There are three different isoforms of the p110 subunit in mammals, p110α, p110β and p110δ, transcribed from the genes *PIK3CA*, *PIK3CB* and *PIK3CD*, respectively, and three isoforms of the p85 subunit, p85α, p55α and p50α, deriving from three genes *PIK3R1*, *PIK3R2* and *PIK3R3*, respectively (Zardavas, Phillips and Loi, 2014).

Signalling through the PI3K class IA pathway begins with the receipt of cell growth and survival signals sensed and relayed to the internal cellular environment by RTKs, such as epidermal growth factor receptor (EGFR), HER2 and insulin–like growth factor-1 (IGF-1), among others, spanning the plasma membrane (Sansal and Sellers, 2004; Keegan *et al.*, 2018). Upon growth factor stimulation, p85 binds to phospho-motifs of RTKs, relieving its inhibitory effect over p110 and mediating the recruitment of PI3K to the plasma membrane. The activated p110 sub-unit catalyses the conversion of PIP2 to PIP3 (Zardavas, Phillips and Loi, 2014). Once generated, the phospholipid PIP3 serves as a nidus for recruiting certain kinases to the plasma membrane, including the Protein kinase B/AKT family of kinases and phosphoinositide-dependent kinase 1 (PDK1). On membrane localization AKT is activated, in part through phosphorylation by PDK1, and is then capable of phosphorylating a number of downstream targets (Sansal and Sellers, 2004). These AKT targets or substrates play key roles in regulating critical cellular functions including proliferation, apoptosis, glucose homeostasis, cell size, nutrient response and DNA damage (Sansal and Sellers, 2004). The signalling outputs of the PI3K pathway, through AKT and other effectors, lead to alterations in multiple cellular processes including cell-cycle regulation, cell-survival, cell adhesion and motility, angiogenesis, glucose homeostasis, and cell size and organ size control (Figure 1.5) (Sansal and Sellers, 2004).

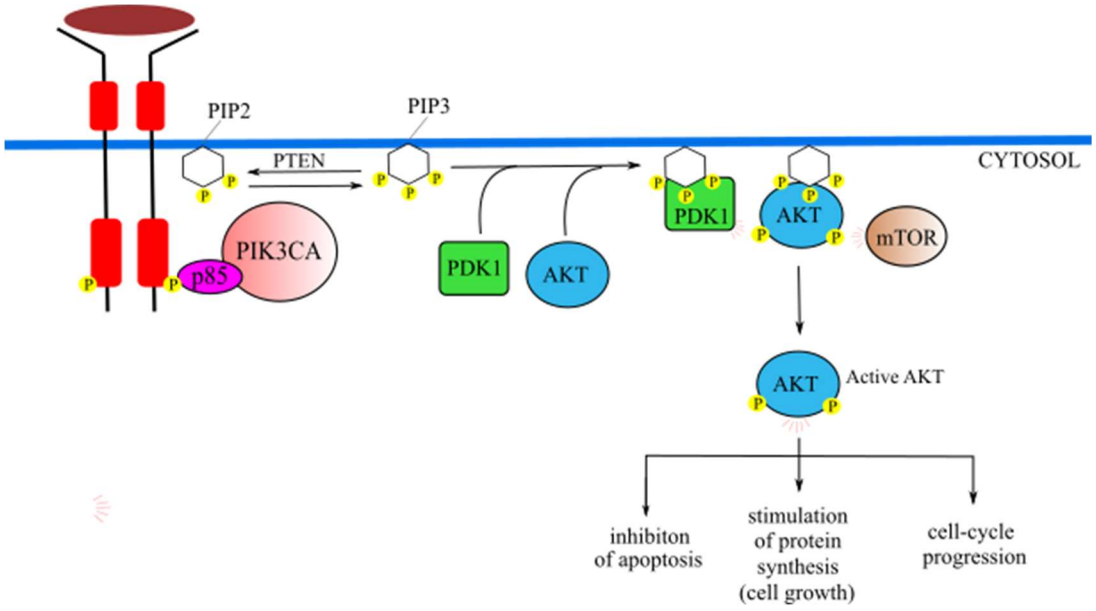


Figure 1.5: Simplified scheme of PI3K/AKT pathway. An extracellular survival signal activates an RTK, which recruits and activates PI 3-kinase. The PI3K produces PI(3,4,5)P3, which serves as a docking site for two serine/threonine kinases—AKT and the phosphoinositide-dependent kinase PDK1—and brings them into proximity at the plasma membrane. The AKT is phosphorylated on a serine by a third kinase (usually mTOR in complex 2), which alters the conformation of the AKT so that it can be phosphorylated on a threonine by PDK1, which activates the AKT. The activated AKT now dissociates from the plasma membrane and phosphorylates various target proteins, promoting inhibition of apoptosis, stimulation of protein synthesis and cell-cycle progression. (Alberts et al., 2014)

PIP3 is negatively regulated by dephosphorylation by phosphatase and tensin homolog (PTEN), which negatively regulates the activation of the PI3K/AKT pathway (Parker *et al.*, 2009). PTEN is a tumour suppressor and acts as a negative control of the PI3K/AKT pathway. Germline and somatic mutations of PTEN, frequently truncating mutations, lead to the disruption of the protein and are implicated in a broad panel of human cancers, including breast cancer (Bachelot *et al.*, 2011).

After the *TP53* gene, the phosphatidylinositol-4,5-bisphosphate 3-kinase, catalytic subunit alpha (*PIK3CA*) gene is the most frequently mutated gene in breast cancer (Pang *et al.*, 2014), with a reported frequency of 20–40% (Cossu-Rocca *et al.*, 2015). The mutations result in PI3K activation independent of upstream signalling and constitutive activation of the downstream AKT pathway, thus contributing to oncogenesis (Pang *et al.*, 2014). *PIK3CA* mutations are more prevalent in ER/PR positive (35%) and HER2 positive breast cancer (23%), than in triple negative breast cancer (ranging from 5% to 13.2%) (Cossu-Rocca *et al.*, 2015), and it is amplified frequently in ER-negative Basal-like breast cancers (Bhat-Nakshatri *et al.*, 2016). Notably, three recurrent oncogenic ‘‘hotspot’’ mutations comprise the majority of somatic *PIK3CA* mutations (Figure 1.6). Two of these mutations, E542K and E545K, occur in

the helical domain found in exon 9, and the third mutation, H1047R, affects the kinase domain located within exon 20 (Wang *et al.*, 2009). The three hotspots represent more than 85% of *PIK3CA* mutations and lead to constitutive PI3K activity by different mechanisms (Pang *et al.*, 2014), with E542K mutation representing 11%, E545K mutation representing 20% and H1047R mutation representing 55% of *PIK3CA* mutations in breast cancer (Bhat-Nakshatri *et al.*, 2016). In vitro and X-ray crystallography data suggest that the E542K/ E545K mutants, located at helical domains, abrogate the intermolecular interaction, whereas the kinase domain mutant, H1047R, promotes constitutive activation of PI3K (Bachelot *et al.*, 2011).

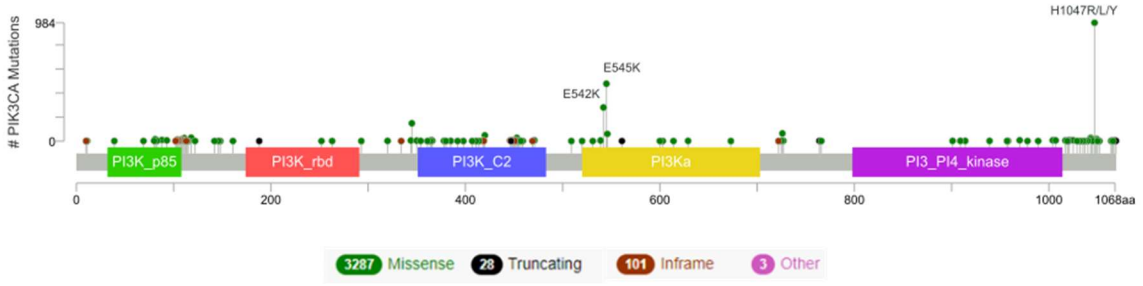


Figure 1.6: p110α's protein amino acid alterations. Schematic lollipop representation of number of *PIK3CA's* protein amino acid alterations from combination of 14 studies from cBioPortal. Diagram circles are coloured with respect to the corresponding type of alteration. In case of different mutation types at a single position, colour of the circle is determined with respect to the most frequent type. PI3K_p85: N-terminal adaptor-binding domain that binds to p85; PI3K_rbd: ras binding domain; PI3K_C2: membrane binding domain; PI3Ka: helical domain; PI3_PI4_kinase: kinase catalytic domain (Adapted from cBioPortal, Cerami *et al.,* 2012; Gao *et al*., 2013).

These mutations join *PIK3CA* amplification, *PTEN* loss, *AKT* mutations and RTK amplification in a class of frequent genomic aberrations that promote tumorigenesis through upregulation of the PI3K/AKT signalling axis (Yuan and Cantley, 2008). These modifications result in the proliferation, survival, and migration of cells, which then result in cell transformation and tumour progression (Wang *et al.*, 2017). Aberrant PI3K signalling can also lead to oestrogen- independent growth of breast cancer cells (Keegan *et al.*, 2018).

While *PIK3CA* mutation is frequent, clinically, the prognostic significance of detecting somatic *PIK3CA* mutation in a breast tumour is uncertain. In a single centre retrospective study of 590 patients with early stage breast cancer, Kalinsky *et al.* found that compared with wild-type, *PIK3CA* mutated tumours were significantly more likely to occur in elderly patients with low grade, lymph node negative, ER positive, HER2 negative breast cancers (Kalinsky *et al.*, 2009). Similarly, in an analysis of 4,294 patients treated in the Tamoxifen Exemestane Adjuvant Multinational (TEAM) phase III trial, *PIK3CA* mutations were associated with lower grade,

fewer involved lymph nodes and progesterone receptor positivity, all considered good prognosis indicators (Sabine *et al.*, 2014). Conversely, a study of 1,394 early breast cancers demonstrated an association between *PIK3CA* mutations and poor prognostic variables (Aleskandarany *et al.*, 2010). Changes in PI3K activity have also been associated with resistance to endocrine therapy, chemotherapy, radiotherapy and anti-HER2 therapy (Keegan *et al.*, 2018).

Given the frequency and effect of the *PIK3CA* mutation, it was hoped that clinically, it would translate to an excellent predictor of response to PI3K inhibitor therapy. However, this has not been fully borne out in clinical studies so far (Press, 2015; Keegan *et al.*, 2018). Clinical trial evidence, thus far, has demonstrated modest efficacy in broad populations based on traditional classifications of ER positive or HER2 positive breast cancer. Phase III trials have reported some significant toxicity with only small gains in progression free survival and no clarity on the question of *PIK3CA* mutation status as a predictor of response (Keegan *et al.*, 2018).

TP53

TP53 gene is mutated in approximately 30% of breast cancers and the prevalence of *TP53* mutations is higher in recurrent tumours than in primary ones. Mutations in the *TP53* gene confer a worse overall and disease-free survival in breast cancer cases, and this effect is independent of other risk factors. In several of the studies the presence of a *TP53* mutation was the single most adverse prognostic indicator for both recurrence and death (Borresen-Dale, 2003). They were associated with aggressive characteristics like high grade, large size, and lymph node positivity and thus related to a poor clinical outcome (Bachelot *et al.*, 2011).

The *TP53* gene is located on chromosome 17p13 and is composed of 19180 bp, spanning 11 exons and 10 introns. The coding sequence starts in the 2nd exon and ends in the 11th, giving rise to p53, a 393-amino-acid protein. This 53 kDa protein can be schematically divided into three main domains: the transactivation domain, the DNA binding domain and the oligomerization domain. Each domain plays an important role in p53 functions. The transactivation domain (TAD), encoded by exons 2 and 3 is serine and threonine-rich and the site of phosphorylation by ATM, ATR or CHK2, which induce protein activation. This first domain also permits the interaction of p53 with numerous proteins that regulate p53's activation, acting as transcription regulators or as transcription factors, such as CBP, CREB, MDM2 and p300. The DNA binding domain is encoded by exons 5 to 8 and recognizes a

consensus sequence in some promoter sequences. Finally, the oligomerization domain is encoded by exons 9 and 10. Thanks to this domain, p53 is able to interact with itself to form an active tetramer (Végran *et al.*, 2013).

In normal cells, p53 protein is maintained at low levels by a series of regulators including MDM2, which functions as a p53 ubiquitin ligase to facilitate its degradation. However, p53 is stabilized in response to various cellular stresses, including DNA damage and replication stress produced by deregulated oncogenes. Mechanisms leading to p53 activation can be stimulus dependent: for example, DNA damage promotes p53 phosphorylation, blocking MDM2-mediated degradation, whereas oncogenic signalling induces the ARF tumour suppressor to inhibit MDM2 (Kastenhuber and Lowe, 2017).

The currently accepted model for the function of wild-type p53 protein is a multifunctional transcription factor involved in the control of cell cycle progression, DNA maintenance and genome integrity, repair after DNA damage, and apoptosis. Indeed, p53 is crucial for a reversible DNA damage-induced G1 phase checkpoint that is mediated, in part, by its ability to transcriptionally activate the p21 cyclin-dependent kinase inhibitor gene, presumably facilitating DNA repair prior to further cell division. Loss of p53 function eliminates the growth arrest response to DNA damage and may allow replication of damaged template DNA. In some circumstances, p53 induces cellular senescence, a stable if not permanent cell cycle arrest program that also involves the retinoblastoma (RB) gene product. p53 can also promote apoptosis, relying on the induction of pro-apoptotic BCL-2 family members whose action facilitates caspase activation and cell death. Why p53 promotes cell cycle arrest in some cell types and apoptosis in others is incompletely understood (Borresen-Dale, 2003; Kastenhuber and Lowe, 2017).

Interestingly, the majority (74% to 81%) of p53 mutations found in human tumors are missense mutations, which usually result in the expression of full-length, albeit mutant, p53 proteins. Although p53 mutations have been found in all coding exons of the p53 gene, the majority of the missense mutations are clustered in the exons that code for the p53 DNA-binding domain, resulting in the loss of DNA-binding activity of mutant p53, which is consistent with the importance of p53-mediated transcription in tumour suppression (Sigal and Rotter, 2000; Liu, Zhang and Feng, 2013; Végran *et al.*, 2013; Kastenhuber and Lowe, 2017; Duffy, Synnott and Crown, 2018). Furthermore, 25% of p53 mutations occur at six ‘mutational hotspots’ in the DNA-binding domain of p53, including residues R175, G245, R248, R249,

R273, and R282 (Figure 1.7). The frequency of mutations varies with the histological and biochemical characteristics of the breast cancers, being more common in ductal than lobular, in lymph node positive than in lymph node negative, in ER negative than in ER positive, and in HER2 positive than in HER2 negative cases (Liu, Zhang and Feng, 2013; Duffy, Synnott and Crown, 2018).

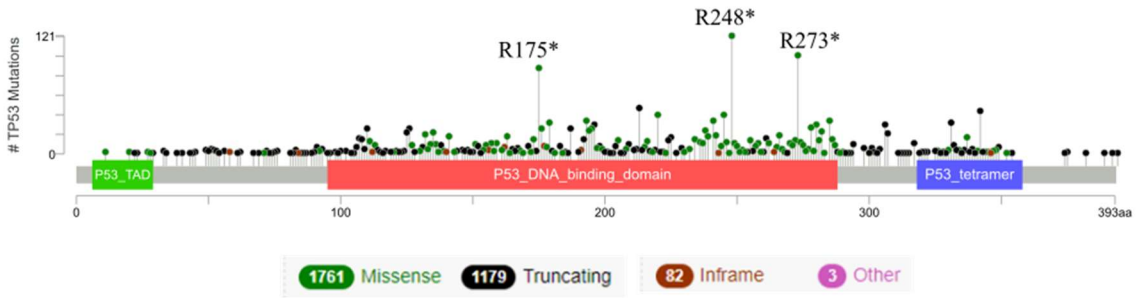


Figure 1.7: p53's protein amino acid alterations. Schematic lollipop representation of number of p53's protein amino acid alterations from combination of 14 studies from cBioPortal. Diagram circles are coloured with respect to the corresponding type of alteration. In case of different types at a single position, colour of the circle is determined with respect to the most frequent type. TAD: transactivation domain (Adapted from cBioPortal, Cerami *et al.,* 2012; Gao *et al*., 2013).

The generally accepted mechanism behind mutant p53 *trans*- dominant suppression (the dominant negative effect) is the shutdown of the wild-type p53 function because of the heteromerization with the mutant p53. Wild-type p53 forms a tetramer to perform its tumor suppressor activity, and this oligomerization is mediated by the oligomerization region (residues 319 –360). This region is fully functional in core domain mutants, allowing complex formations between mutant and wild-type, thus driving wild-type p53 into a mutant or perhaps inactive conformation (Sigal and Rotter, 2000).

Wild-type p53 clearly acts as a negative regulator of cell growth but considering *TP53* mutations solely as loss-of-function mutations would prevent a full understanding of how *TP53* mutations drive tumor growth. The status of this gene in human cancer is often defined in binary terms, wild-type versus inactivated, despite accumulating evidence that the majority of mutant p53 proteins are heterogeneous oncogenic proteins with multiple gain-of-function (GOF) activities and with potential as therapeutic targets. Mutant p53 GOF activities include enhanced tumorigenesis, metastasis, resistance to therapy and genomic instability. GOF mechanisms may be the result of changes in the specificity of the DNA-binding activity of the p53 mutant,

leading to the induction of novel transcriptional programs, or changes on its interaction with other cellular proteins, directly or indirectly related to the regulation of gene expression (Soussi and Wiman, 2015).

TP53 mutations are not the only way to inactivate its tumour suppressor function. Several other mechanisms for inactivation of p53 itself have been described including amplification of one of the many p53 binding proteins such as MDM2, alterations in genes coding for proteins responsible for the phosphorylation, acetylation and ribosylation of the p53 protein, like ATM and CHK2, and in genes coding for transcription factors of the *TP53* gene itself, like *HoxA5* (Borresen-Dale, 2003).

While a biomarker potential for mutant p53 has been widely investigated in breast cancer, measurement of the mutant protein has not been validated for clinical use. In contrast to a biomarker role, little work until recently had been done on exploiting the mutant protein as a target to treat breast cancer. Molecules and approaches aiming to reactivate mutant p53 or activate wild-type p53 are being developed and some of them are in phase I trials (Bachelot *et al.*, 2011; Duffy, Synnott and Crown, 2018).

Indeed, the difficulties associated with exploiting p53 therapeutically do not mitigate the astounding morbidity associated with *TP53* mutation. In the absence of new therapeutic innovations, *TP53* mutant cancer will lead to the deaths of more than 500 million people alive today. New technologies, together with the ever-increasing understanding of the complexity of p53 action and the diverse consequences of p53 mutations, will hopefully set the stage for more robust clinical advances (Kastenhuber and Lowe, 2017).

1.2 Differential Expression of Mutations

Despite the knowledge coming from the extensive cataloguing of mutations in cancer, these mutations alone are not enough to correctly predict outcome and drive the best management of breast cancer patients. In this way, the scientific community is still trying to find new and better biomarkers that will refine the molecular classification of breast cancer and provide new prognostic factors and new targets for therapy (Kim *et al.*, 2004; Santarpia *et al.*, 2016). Another way to study the heterogeneity of breast cancers is through the analysis of differential expression of mutations and the mechanisms driving it. In this section I expand the current knowledge on the two main mechanisms: copy number alterations and cis-regulation.

1.2.1 Copy Number Alterations

1.2.1.1 Mutant Allele Specific Imbalance

While activating somatic mutations in one allele of an oncogene (heterozygous mutation, ‘‘one hit’’) are generally believed sufficient to confer a selective growth advantage on the cell, mutant allele specific imbalance (MASI) has been observed in tumours and cell lines harbouring oncogenic mutations (Soh *et al.*, 2009). In rare instances, the mutant allele becomes dominant, either through deletion of the wild-type allele and/or through copy number gain of the mutant allele. This phenomenon was termed mutant allele-specific imbalance (MASI) and reflects increased copy number dosage of the mutant allele (Krasinskas *et al.*, 2013).

In 1991, Mitsudomi *et al.* first described differential expression of *KRAS* mutations in non-small-cell lung cancer (Mitsudomi *et al.*, 1991). In a study of pancreatic, lung and colorectal cancers, mutant allele specific imbalance of *EGFR* and *KRAS* mutations have been reported (Soh *et al.*, 2009). Predominant expression of the *KRAS* mutant allele has also been associated with worse clinical outcome in lung (Chiosea *et al.*, 2011), colorectal (Hartman *et al.*, 2012; Bielski *et al.*, 2018) and pancreas (Krasinskas *et al.*, 2013; Bielski *et al.*, 2018) carcinomas and with resistance to treatment with cetuximab (anti-EGFR antibody) in an in vitro model of metastatic colorectal cancer (Malapelle *et al.*, 2015). Also, *EGFR* MASI in lung adenocarcinomas could be used to identify younger patients with more aggressive disease (Oakley and Chiosea, 2011).

Another MASI observation is that patients with melanoma displaying *BRAF* mutant allele percentage similar to or higher than that of the wild-type allele seemed to benefit the most from MAPK inhibitors therapy and patients with balanced heterozygosity had longer progression free survival than those with allelic imbalance (Stagni *et al.*, 2018). A recent study showed that a subset of patients with loss of wild-type *BRAF* had markedly improved progression-free survival compared with patients that were either heterozygous or possessed genomic gains of the V600E allele (Bielski *et al.*, 2018).

As observed in many studies, cancer cells gain a selective growth advantage by tuning the dosage and stoichiometry of oncogenic driver mutations in a gene- and lineage-specific manner, driven in part by competitive fitness, alluding to a broader growth-suppressive effect of the WT allele on mutant oncogenes. These data suggest that the presence and type of oncogenic driver

mutations allelic imbalance among therapeutically actionable oncogenes may provide novel biomarkers of drug sensitivity with broad implications for tumour biology and precision oncology (Bielski *et al.*, 2018).

1.2.1.2 Integrative Clusters

Recently, performing an integrated analysis of somatic copy number aberrations (CNAs) and gene expression profiles in 2,000 primary breast cancers from the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC), Curtis *et al.* classified these tumours in ten integrative cluster (IntClust) subtypes with distinctive copy number profiles and clinical courses (Curtis *et al.*, 2012). To date, this classification represents the most extensive molecular-based taxonomy of breast cancer. It partly captures subgroups defined by other approaches, but importantly also groups tumours in more novel subtypes, with a distinct clinical outcome (Russnes *et al.*, 2017).

One of the key features of the IntClust classification is stratification of ER positive tumours. There are seven subtypes (IntClusts 1, 2, 3, 6, 7, 8 and 9) that are predominantly composed of ER positive/HER2 negative breast cancers, including one ER positive group (IntClust 2) with poor prognosis (Pereira *et al.*, 2016). From the remaining subtypes, IntClust 5 is composed of HER2 positive tumours; triple negative breast cancers predominantly fall in IntClust 10 with some in IntClust 4, which is characterized by immune cell infiltration (Provenzano, Ulaner and Chin, 2018).

In an analysis based on a multistate statistical model, that yields individual risk-of-relapse estimates using tumour features (clinical, pathological and molecular covariates, and disease chronology), Rueda *et al.* showed that important differences in recurrence rates obscured in the immunohistochemistry and PAM50 subtypes, became apparent when breast tumours were classified into the integrative subtypes. The probabilities of distant recurrence or cancer-related death among triple negative breast cancer patients who were disease-free at five years after diagnosis revealed low (IntClust 10) and high (IntClust4 ER negative) risks for late-relapse, whereas IHC (and PAM50) subtypes homogenized this risk. Marked differences were also apparent among ER positive patients, with patients with IntClust 3, IntClust 7, IntClust 8 and IntClust 4 ER positive subtypes exhibiting a better prognosis, whereas patients with IntClust 1, IntClust 2, IntClust 6 and IntClust 9 subtypes exhibited late-recurring cancer with a poor

prognosis. The IntClust 2 (ER+ tumours) subtype exhibited the worst prognosis, with a probability of relapse second only to that of IntClust 5 (HER2 positive tumours). Collectively, these subgroups comprise 26% of ER positive cases, and thus define the minority of patients who may benefit from extended monitoring and treatment given the chronic nature of their disease. Information about the dynamics of late relapse that is provided by integrative subtype could not be inferred from standard clinical variables, including immunohistochemistry subtype (Rueda *et al.*, 2019).

1.2.2 Cis-regulation of Gene Expression

Although there is evidence of the influence of allelic imbalance in tumour progression and outcome, all studies on mutant allele imbalance have examined the clinical association of somatic mutations' MASI solely with copy number gain at the DNA level. Also, the new IntClust classification only takes into account mutations and copy number alterations. However, cis-regulation of gene expression, particularly by genetic variants, is another mechanism capable of modulating gene expression dosage, and is yet unexplored (Figure 1.8).

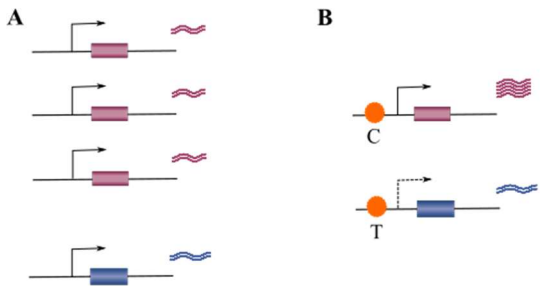


Figure 1.8: Differential Expression of alleles. Purple allele is more expressed than blue allele in (A) due to copy number alteration at DNA level, and in (B) due to cis-regulation.

Gene expression is a complex trait that is influenced by cis- and trans-acting genetic and epigenetic variation and by environmental factors (Pastinen, 2010). Trans-acting regulatory elements, such as transcription factors, regulate both alleles of the gene equally and can be located on the same or on different chromosomes. Cis-acting variants (such as DNA polymorphisms and methylation) are commonly thought to involve regulatory elements such as promoters and enhancers, which may lie immediately upstream of the gene, but can also be found hundreds of kilobases away, and regulate only one of the alleles of a gene (Pastinen and Hudson, 2004; Pastinen, Ge and Hudson, 2006).

Polymorphisms in cis-regulatory sequences can lead to differences in levels of expression between the two alleles that can be extreme (greater than ten-fold difference) or can be more subtle. Even subtle expression differences are still potentially important as a mechanism that has an impact on genotype–phenotype correlation (Chess, 2012).

The most common form of polymorphism is a single nucleotide polymorphism (SNP) (Alberts *et al.*, 2014). SNP is a difference between chromosomes in the base present at a particular site in the DNA sequence. For example, some chromosomes in a population may have a C at that site (the 'C allele'), whereas others have a T (the 'T allele). It has been estimated that, in the world's human population, about 10 million sites (that is one variant per 300 bases on average) vary such that both alleles are observed in a frequency of ≥ 1%, and these 10 million common SNPs constitute 90% of the variation in the population. The remaining 10% is due to a vast array of variants that are each rare in the population (The International Hapmap Consortium, 2003).

SNPs may alter the activity of cis-regulatory elements by affecting protein binding to the DNA (at promoters, enhancers, insulators, and silencers, thus affecting transcription initiation), protein binding to RNA (altering splicing) and miRNA binding (affecting both pre-miRNA sequence, recognition of target) (Pastinen, Ge and Hudson, 2006).

SNPs can also occur within genes and, thus an organism may carry two different alleles of a gene, in which case, with respect to this allele, it is said to be heterozygous. Conversely, the presence of two identical alleles of a gene in an organism's genome renders this organism homozygous with respect to this allele. An allele that is present in the great majority of individuals within a species is usually termed wild-type, the term implying that such an allele, being naturally present in large numbers of apparently healthy organisms, is compatible with normal structure and function (Weinberg, 2013).

The previous assumption is that each allele of a gene was expressed at a similar intensity (Benitez, Cheng and Deng, 2017). However, allelic differences in expression have been observed in familial and population studies. In the most extreme form of differential allelic expression, monoallelic expression, only one of the two alleles in heterozygotes is expressed. Until recently, it was thought that monoallelic expression occurred only for a few kinds of genes: those that undergo allelic exclusion (e.g., immunoglobulins, odorant receptors) or imprinting and X-linked genes affected by the inactivation of one X chromosome in females. Recently, investigators used multiple clones of cells from the same heterozygous individual and concluded that ''monoallelic'' expression is found for 5%–10% of autosomal genes in a large proportion of the clones studied (Cheung *et al.*, 2008).

Direct assessment of cis-regulatory variation requires allele-specific approaches, such as differential allelic expression (DAE) analysis (Pastinen, 2010). DAE can be calculated as the relative expression of two alleles for individuals who are heterozygous at a particular SNP that is present in the mRNA, and distortions in the expected 1:1 ratio of allele A to allele B in the RNA signal can be an indication of allelic imbalance (Liu *et al.*, 2012). DAE assays are optimal for detecting cis-acting differences, as each allele serves as an internal control for the other, and both alleles are impacted by the same trans-acting and environmental effects. Thus, DAE mapping reduces the complexity of gene expression to its cis components, being the most robust approach to assess allelic imbalances. By measuring the ratio of alleles, even subtle cis-acting differences can be revealed (Pastinen, Ge and Hudson, 2006; Adoue *et al.*, 2014).

Preliminary data and Hypothesis

Previously, Maia et al. showed that DAE in normal breast tissue was common for a set of candidate genes associated with breast cancer, such as *BRCA1*, *BRCA2*, *CCND3*, *EMSY*, *GPX4*, *TRXR2*, *MTHFR* and *TP53* (Maia *et al.*, 2009, 2012). Preliminary data from Professor Ana Teresa Maia's group have also shown *PIK3CA* being regulated by *cis*-regulatory variation in normal breast tissue and identified rs2699887 as a functional regulatory SNP, by possibly affecting the binding of NF-YA in the promoter region of *PIK3CA* (manuscript under preparation). Also, there is data on other genes of the same pathway, such as *EGFR, AKT1, TSC1/2, KRAS, BRAF, MAPK1, MAP2K, FRAP, RPS6KB1/2* and *EIF4EBP1*, being cis-regulated in normal breast tissue. (Unpublished data)

Even though there is evidence of DAE due to cis-regulation of genes commonly associated with breast cancer in normal breast tissue, there is no study on DAE of somatic mutations of breast cancer and whether there is any clinical difference between tumours that express more of the mutated allele, tumours with equal amounts of mutated and wild-type alleles and tumours that express more of the wild-type allele.

So, we hypothesise that DAE of mutated genes contributes to breast cancer biology and outcome, by creating imbalances in the allelic levels of gene expression of somatic mutations. Namely, somatic mutations in oncogenes and tumour suppressor genes whose expression is modelled by cis-regulatory variation will have different impact on biology and clinical behaviour of the tumour that harbours them.

To test this hypothesis, the impact of differential expression of *PIK3CA*'s and *TP53*'s somatic mutations on tumour biology and patients survival were assessed using data from two independent sets of breast tumours, the METABRIC (Curtis *et al.*, 2012) and the TCGA projects (Koboldt, Fulton and McLellan, 2012). The Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) and The Cancer Genome Atlas (TCGA) project are two large cohorts for which high-resolution molecular analyses at multiple levels of breast cancer samples have been performed and for which there is also clinical data available.

2 Aims

The main aim of this study is to assess the influence of differential allelic imbalance in breast cancer clinicopathological characteristics.

In order to achieve this objective, the following specific aims are proposed:

1. Select of the two most mutated genes in two different large-scale projects, METABRIC and TCGA
2. Calculate mutant allelic ratios using DNA sequencing (DNAseq) and RNA sequencing (RNAseq) data from METABRIC and TCGA projects
3. Correlate the mutant allelic ratios of the two genes with clinicopathological characteristics of breast cancer and with patients' survival

CHAPTER 3
MATERIALS AND METHODS

3 Material and Methods

3.1 R Programming Language

R is a language and environment that provides a wide variety of statistical and graphical techniques and is highly extensible. The capabilities of R are extended through user-created *packages*, which allow specialised statistical techniques, graphical devices and reporting tools. These packages include reusable R functions and the documentation that describes how to use them (R Core Team, 2018). All the analyses were done in RStudio which is an integrated development environment (IDE) for R. It includes a console, syntax-highlighting editor that supports direct code execution, as well as tools for plotting, history, debugging and workspace management (https://www.rstudio.com/products/rstudio/), making it easier to work with R.

Part of the time dedicated to the present dissertation was spent learning R programming and package implementation.

3.2 Breast Cancer Samples

For the analysis, samples from two different projects were used. The METABRIC set of tumour samples included 2433 breast cancer samples from the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) project (Curtis *et al.*, 2012) with DNA sequencing (DNAseq) data, among which 431 were subjected to a capture-based RNA sequencing (RNAseq) study. The TCGA set was comprised of 695 breast cancer samples from The Cancer Genome Atlas (TCGA) with DNA and RNA sequencing data. The tumours were primary invasive ductal carcinomas of the breast and the samples were collected before any treatment. Both the METABRIC and TCGA sets had been previously pre-processed by our collaborators at the Cambridge Institute – CRUK, in Cambridge (United Kingdom). Heterozygous genotypes had been called from DNA data to avoid RNA editing confounding, and other RNA related variants interference, and because true allelic imbalance can lead to heterozygous sites being called as homozygous in RNA-based genotype calling.

For the *PIK3CA's* analysis from METABRIC project, I received a file with a subset of the data with only DNAseq and RNAseq data for samples with *PIK3CA* mutations. Then, I received a second file with all RNAseq data from the METABRIC project. Since this second file had been pre-processed a few years earlier, before the beginning of the analysis of this

second file, I manually curated the Catalogue of Somatic Mutations in Cancer (COSMIC) classification of the mutations, in order to exclude mutations that were re-classified as normal variants (SNPs) in the population after the catalogue was updated. I manually checked, from April 2nd to April 5th, 2019, 14250 observations that had COSMIC ID on the COSMIC website (https://cancer.sanger.ac.uk/cosmic). All rows for which the COSMIC ID was reclassified as SNP, were excluded for the remainder of the analysis, leading to a total of 8749 observations to further analyse. This extra step for the METABRIC set was needed, because I observed that some of the mutations' COSMIC IDs were now classified as SNP, so I decided to check all the other entries as well. DNAseq data for the whole METABRIC set was then obtained from the supplementary material of Pereira *et al.* paper (Pereira *et al.*, 2016) on 24/04/2019.

Data for cellularity analysis of METABRIC set was also downloaded on 19/03/2019 from the supplementary material of Pereira *et al.* paper (Pereira *et al.*, 2016). Tumour cellularity was classified by pathologists in high, medium and low levels. High cellularity was set at more than 70% of tumour cells in the sample (i.e. tumours with higher purity), medium cellularity was set at 40-70% of tumours cells in the sample, while low cellularity was set at less than 40% of tumours cell in the sample.

3.3 Selection of genes for the analysis

The number of mutations *per* gene on the two datasets were assessed and *PIK3CA* and *TP53,* the two genes with higher number of mutations, were chosen for further analysis. Two separate analysis were then performed, one assessing the allelic imbalance of *PIK3CA* somatic missense mutations and another one assessing the allelic imbalance of *TP53* somatic mutations.

3.4 Statistical Analysis of Allelic Expression Imbalances

Prior to the analysis, a set of filtering steps were performed. Firstly, samples which passed quality control were selected in order to exclude samples which had low quality of sequencing information. Then, samples were filtered in order to keep only those harbouring exonic somatic mutations. Since *PIK3CA* is an oncogene, only missense mutations were selected for its analysis (Pang *et al.*, 2014), to simplify the analysis. For the *TP53* analysis, all types of mutations were analysed without discrimination, since *TP53* has a dual role, sometimes acting as a tumour suppressor gene and sometimes as an oncogene (Soussi and Wiman, 2015).

Next, samples with less than 30 reads for both RNAseq and DNAseq data were excluded, which has been used in our group in other studies. Lastly, the samples with more than one mutation were identified, and data for the mutation with the higher number of reads for RNAseq was kept. When the samples had the same depth for RNAseq, the DNAseq's highest depth was used to choose the sample to keep for further analysis. The exclusion of duplicated samples has been done to simplify the analysis. Clinical information was then obtained from cBioPortal for TCGA set (Cerami *et al.*, 2012; Gao *et al.*, 2013) while clinical information for METABRIC set was provided by our collaborators at the Cambridge Institute – CRUK, in Cambridge (United Kingdom).

Demographic features and disease characteristics of the two datasets after filtering process are summarized in Annex B and Annex O for *PIK3CA*'s and *TP53*'s analyses, respectively. The workflow of the analysis is displayed in Figure 3.1.

For all samples three parameters were calculated: 1) The mutant allele expression ratio $\alpha = Log_2[(\text{mutant allele count in RNA})/(\text{wild-type allele count in RNA})]$, that served as a measure of the net allelic expression imbalance in the tumours; 2) The DNA mutant allele ratio $\beta = Log_2[(\text{mutant allele count in DNA})/(\text{wild-type allele count in DNA})]$, which served to control for sequencing artefacts from heterozygous genotypes and to account for differences in variant frequencies in DNA; and 3) the normalized mutant allele expression ratio $\gamma = [\alpha]–[\beta]$, which is a measure of mutant allelic expression imbalance due to *cis*-regulation alone. Since the parameters calculated are fractions, the denominator cannot have the value zero, because its overall value would be undefined. So, samples with wild-type allele count equal zero had to be excluded from the analysis.

The choice of using relative expression of the alleles to study differential allelic expression is based on the knowledge that the elimination of environmental or trans- acting influences that alter gene expression or DNA–protein interactions should provide higher sensitivity for uncovering the direct influence of sequence and epigenetic variation in cis-regulatory elements (Pastinen, 2010).

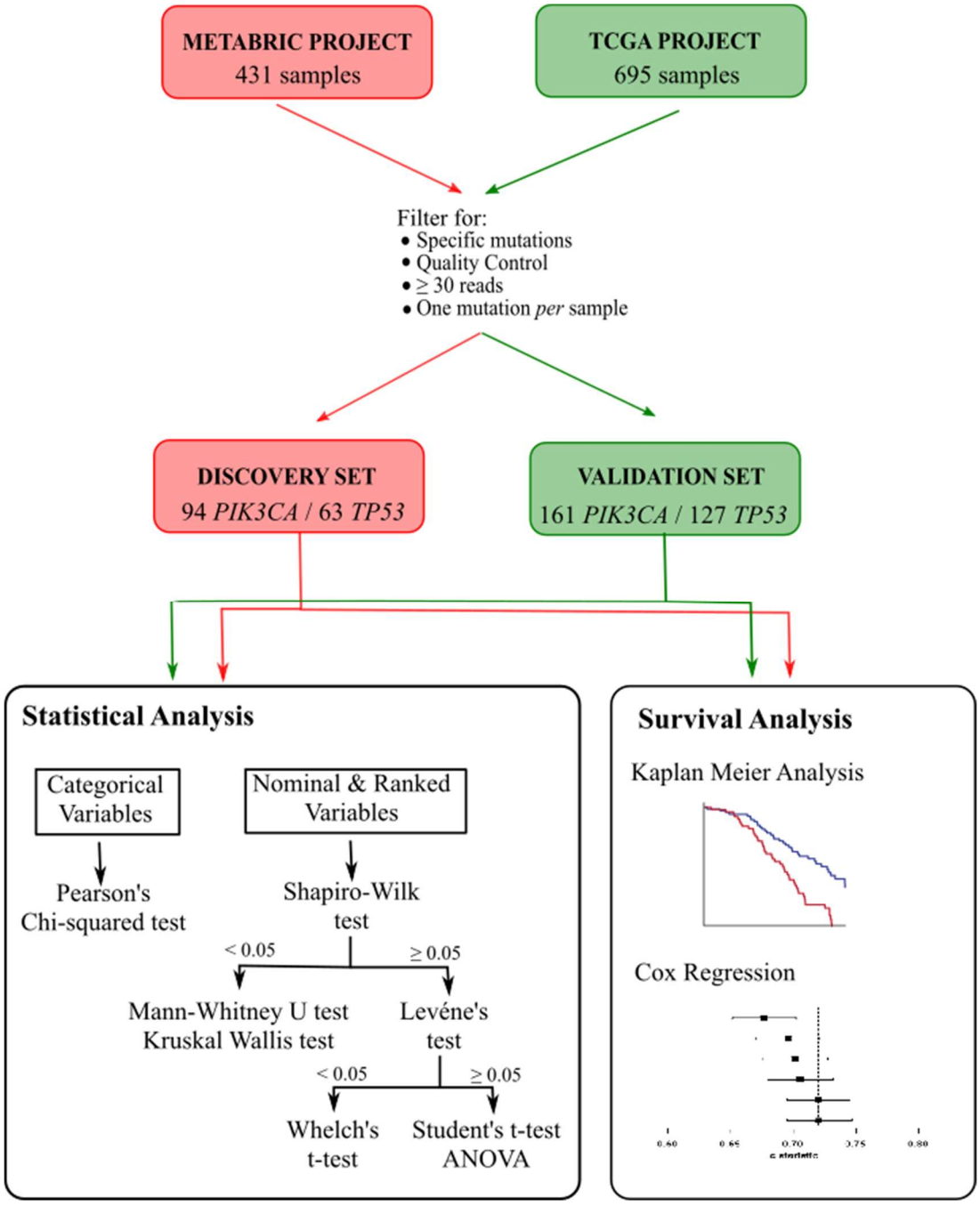


Figure 3.1: Flow chart of *PIK3CA's* and *TP53's* Analysis. First the METABRIC dataset was filtered to keep only unique samples of *PIK3CA's* missense or *TP53's* somatic mutations that passed quality control and had more or equal 30 reads. After the filtering step, the METABRIC set which was comprised of 94 samples and of 63 samples for PIK3CA and TP53 analysis, respectively, was analysed by statistical and survival analyses. When analysing associations between categorical variables, Chi-squared test was applied. If the association to be studied was between nominal and ranked variables, Shapiro-Wilk test was first applied and if statistically significant, Mann-Whitney U test or Kruskal-Wallis test were applied. If not significant, Levéne's test was applied and next Whelch's t-test or Student's t-test were applied, depending on Levéne's being significant or not significant, respectively. The survival analysis was comprised of Kaplan-Meier and Cox Regression Analysis. Next, the TCGA's 695 breast cancer samples were subjected to the same filtering steps as described above. Then, the same analysis described above was made for the TCGA set (161 samples and 127 samples, for *PIK3CA* and *TP53* respectively). All the analyses described were done in R.

Additionally, we considered that there was mutant allele differential expression (MADE) when $|\gamma| \geq 0.58$ (1.5 fold or greater difference). If $-0.58 < \gamma < 0.58$, we considered that the expressions of mutated and wild-type alleles were not different (mut=wt). For the survival analysis, we further separated the MADE samples into MADE_wt ($\gamma \leq -0.58$, or preferential expression of wild-type allele) and MADE_mut ($\gamma \geq 0.58$, or preferential expression of mutant allele). The threshold |0.58| is arbitrary and determined by the sensitivity and specificity of the applied differential allelic expression detection method, as used previously (Maia *et al.*, 2009).

The α ratio was further divided into groups, one with samples with no differential allelic expression ($|\alpha| \leq 0.58$, α_noDAE group) and another with samples with differential allelic expression ($|\alpha| \geq 0.58$, α_DAE group). The α_DAE group, was still sub-divided into α_DAEwt (preferential expression of wild-type allele, $\alpha \leq -0.58$) and α_DAEmut (preferential expression of mutated allele, $\alpha \geq 0.58$) groups.

Association between allelic expression imbalance ratios and clinical data was achieved by bivariate analysis. Whenever the variables were categorical, Pearson's Chi-squared test was used to assess whether the proportions for one variable were different among values of the other variable. The null hypothesis for this test is that the proportions at one variable are the same for different values of the second variable. The observed frequencies are used to calculate the expected frequencies. Once the expected numbers are found, they are compared to the observed numbers using the chi-square test, checking if there is any statistically significant difference (McDonald, 2014).

In order to assess associations with age, lymph node status and size of tumour, the data was binned. Age was divided in <45 years, 45-54 years, 55-64 years and >65 years. Lymph node status was divided into patients with no positive lymph node (0), 1 positive lymph node (1), 2 to 4 positive lymph nodes (2-4), 5 to 9 positive lymph nodes (5-9) and more than 10 positive lymph nodes (10+). Tumour's size was divided in 0-9mm, 10-19mm, 20-29mm, 30-39mm and >40mm in size.

When studying association between nominal and ranked variables, Shapiro-Wilk test was applied first to assess if samples had normal distribution. The normal distribution is the familiar bell-shaped curve, with a high probability of getting an observation near the middle and lower probabilities as you get further from the middle. In this test, the null hypothesis is that the distribution is normal, so if the p-value is significant, we can refute the null hypothesis

and assume the distribution is not normal (McDonald, 2014). The normality of distribution was also checked with the variables being plotted to observe the shape of the curve.

For a normal distributed sample, parametric test like Student's t-test can be applied to test for differences in the mean of two groups. Since parametric tests assume that data are homoscedastic (have the same variance in different groups), Levéne's test was applied to check this assumption. This test's null hypothesis is that the variances are equal, so if the test is not statistically significant, we assume the null could be true, and applied Student's t-test. Student's t-test is used for two samples when one is a measurement variable and the other a nominal variable, and the nominal variable has only two values. It tests whether the means of the measurement variable are different in the two groups. The statistical null hypothesis is that the means of the measurement variable are equal for the two categories. So, if the test is statistically significant, the means can be assumed different. Whenever the nominal variable had more than two values, the Analysis of Variance (ANOVA) was applied, since it generalizes the t-test beyond two means. Welch's t-test was applied when Levéne's test was statistically significant, since it is designed for testing between groups with unequal variances. Welch's t-test is also used to test the hypothesis that populations' means are equal and it can be generalized to more than two samples (McDonald, 2014).

When samples were not normally distributed, Mann-Whitney U tests (also called the Mann–Whitney–Wilcoxon (MWW), Wilcoxon rank-sum test, or Wilcoxon–Mann–Whitney test) was applied. This is the non-parametric analogue to the *t*–test with the null hypothesis that it is equally likely that a randomly selected value from one sample will be less than or greater than a randomly selected value from a second sample. Since the Mann-Whitney U test only accommodates the comparison between two groups, whenever there were more than two groups the Kruskal-Wallis test was used. Kruskal-Wallis test is a non-parametric method, equivalent to ANOVA (McDonald, 2014).

P-values were considered significant when smaller than 0.05. Whenever multiple tests were applied, p-values were adjusted using Bonferroni correction. The adjustment of p-values is done because when a large number of statistical tests are performed, some will have p-values less than 0.05 purely by chance, even if all null hypotheses are really true (Error Type I). The classic approach to the multiple comparison problem is to control the familywise error rate. The familywise error rate is the probability of coming to at least one false conclusion in a series of hypothesis test. The term "familywise" error rate comes from *family of tests*, which is the

technical definition for a series of tests on data. Instead of setting the critical *P* level for significance to 0.05, a lower critical value is used. If all the null hypotheses are true, the probability that the family of tests includes one or more false positives due to chance is 0.05. The Bonferroni correction is the most common way to control the familywise error rate. The critical value for an individual test can be found by dividing the familywise error rate (usually 0.05) by the number of tests (n), i.e. error rate/n. Thus, if doing 100 statistical tests, the critical value for an individual test would be 0.05/100=0.0005, and only individual tests with p-value <0.0005 should be considered significant. The Bonferroni correction is mainly useful when there are a fairly small number of multiple comparisons and only one or two might be significant. An important issue with the Bonferroni correction is deciding what a "family" of statistical tests is. There is no firm rule on this, and individual judgement should be used, based on just how bad a false positive would be. This decision should be made before looking at the results, otherwise it would be too easy to subconsciously rationalize a family size that gives the wanted results. The goal of multiple comparisons corrections is to reduce the number of false positives. An unfortunate by-product of correcting for multiple comparisons is that it may increase the number of false negatives (Error Type II), where the null hypothesis is not rejected erroneously (Bush and Moore, 2012).

3.5 Consequence of Mutations

The classification of mutations according to their consequences was determined using OncoKB, a curated knowledgebase of the oncogenic effects and treatment implications of mutations and cancer genes (http://www.oncokb.org/) (Chakravarty *et al.*, 2017). Each mutation was manually verified on the OncoKB website.

3.6 Genotype analysis of rs2699887

For the analysis performed with the genotypes of rs2699887, I used genotype data from the METABRIC project provided by our collaborators at the Cambridge Institute – CRUK, in Cambridge (United Kingdom). Since rs2699887 was not genotyped in the array used in the METABRIC project, I used the information of the proxy SNP rs2699905, which is in complete

linkage disequilibrium (LD) with the rs2699887 (the analysis of which SNP was in LD with rs2699887 was performed by Dr Joana Xavier).

LD is a non-random association of alleles at different loci on a haplotype in a given population, serving as a measure of the degree to which SNPs at two loci are associated (Schaid, Chen and Larson, 2018). SNPs are in complete LD with each other when they are inherited together. This strong association between SNPs in a region has a practical value: genotyping only a few, carefully chosen SNPs in the region will provide enough information to predict much of the information about the remainder of the common SNPs in that region. As a result, only a few of these 'tag' SNPs are required to identify each of the common haplotypes in a region (Cheung and Spielman, 2009). The fact that the rs2699905 is in complete LD with the SNP identified as regulatory (rs2699887) means that we can use the first SNP as a marker of the second, since there are only two possibilities: whenever rs2699905 is a T allele, rs2699887 is a T allele, and whenever rs2699905 is a C allele, rs2699887 is a C allele. So, by knowing the genotype of rs2699905, we also know the genotype of rs2699887. The rs2699887 T allele is the minor allele, which means it is less frequent in the population, while the C allele is the major allele, which means it is the most frequent allele in the population.

3.7 Survival Analysis

Survival analysis is a collection of statistical procedures for data analysis where the variable of interest is time until an event occurs (Clark *et al.*, 2003). It involves the consideration of the time between a fixed starting point (e.g. diagnosis of cancer) and an outcome (e.g. death). The key feature that distinguishes such data from other types is that the event will not necessarily have occurred in all individuals by the time the study ends, and for these patients, their full survival times are unknown (Bradburn *et al.*, 2003a). This phenomenon is called censoring and it may arise in the following ways: (a) a patient has not (yet) experienced the relevant outcome, such as relapse or death, by the time of the close of the study; (b) a patient is lost to follow-up during the study period; (c) a patient experiences a different event that makes further follow-up impossible. Basically, censored patients are included in estimates of survival probabilities at time points preceding their censoring time point and excluded from the analysis thereafter. Unbiased inferences require the censoring to be non-informative, with the time of censoring absolutely not related to the event time. These time-to-event outcomes offer more information than simply whether or not an event occurred, and to deal with the censored data, the survival techniques were developed (George, Seals and Aban, 2014). The power of a

method to analyze survival time data depends on the number of events rather than the total sample size (Schober and Vetter, 2018).

Time-to-event studies typically employ two closely related statistical approaches, Kaplan-Meier analysis and Cox proportional hazards model analysis (sometimes abbreviated as proportional hazards model or Cox model) (Dudley, Wickham and Coombs, 2016).

In comparing the survival distributions of two or more groups, Kaplan-Meier estimation, which is a univariate analysis, and the log-rank test are the basic statistical methods of analyses. These are non-parametric methods in that no mathematical form of the survival distributions is assumed (George, Seals and Aban, 2014).

The Kaplan-Meier method uses the survival function to estimate the unadjusted probability of surviving past a specified time point, or more generally, the probability that the event of interest has not yet occurred by this time point. A Kaplan-Meier curve shows the estimated survival function by plotting estimated survival probabilities against time. The estimated survival probability is constant between the events. Therefore, the curve is a step-function in which each vertical drop indicates the occurrence of one or more events. The survival curve is unchanged at the time of a censored observation, but at the next event after the censored observation the number of people "at risk" is reduced by the number censored between the two events. The censoring of patients is typically indicated by a vertical mark at the censoring time. The median survival time, which is the time when the event has occurred in 50% of the patients or study subjects, is a commonly reported summary statistic for survival time data (Schober and Vetter, 2018).

The log-rank test is based on a comparison of the observed and expected numbers of events, with the expected number calculated under the assumption of no difference in survival between groups. The log-rank test statistics is compared with the critical values of a Chi-square distribution. The larger the discrepancies between the observed and expected number of events in the groups, the more likely the log-rank test statistics will be larger than the critical value of a Chi-square distribution with one degree of freedom (in the case of two groups); and therefore, more likely to reject the null hypothesis of no difference (Bewick, Cheek and Ball, 2004).

It is well known that the log-rank test is optimal when the two hazard rates are proportional. However, when the survival curves of different groups cross—indicating that one

group has a more favorable survival in a certain time interval and less favorable survival in another time interval and, thus the assumption of proportional hazard rates is violated —the power to detect such differences is very low (Qiu and Sheng, 2008; Schober and Vetter, 2018).

In the literature, there are a fair number of statistical methodologies for addressing the crossing-curves problem. However, in many applications, it is difficult to specify the types of survival differences and choose an appropriate method prior to analysis. In an extensive series of Monte Carlo simulations to investigate the power and type I error rate of these procedures under various patterns of crossing survival curves with different censoring rates and distribution parameters, Li *et al.* demonstrated that the Two-Stage procedure and the adaptive Neyman's smooth tests offered higher power and greater stability than other methods when the survival distributions cross at early, middle or late times. Even for proportional hazards, both methods maintain acceptable power compared with the log-rank test. Because the Two-Stage test is not subject to weight functions, it is a particularly convenient method for addressing all possible alternatives. (Li *et al.*, 2015).

In the first stage of the Two-Stage test, a conventional procedure such as the log-rank test is applied, to detect all kinds of differences between the two hazard rates, except certain crossings. In the second stage, a procedure for detecting crossings is applied. If the null hypothesis is rejected in stage one, then the entire procedure ends, and it can be concluded that the two hazard rates are significantly different. Otherwise, stage two is performed, by applying a procedure for handling the crossing hazard rates problem, from which cases when the two hazard rates are identical can be distinguished from cases when they cross each other (Figure 3.2) (Qiu and Sheng, 2008).

Moreover, the log-rank test and two stage procedure cannot adjust for other covariates that might affect survival time. While it can determine whether observed differences are significant, it cannot provide an estimate of the difference between groups (Schober and Vetter, 2018). In order to adjust for other covariates, a multivariate analysis, such as Cox proportional hazard model, is necessary.

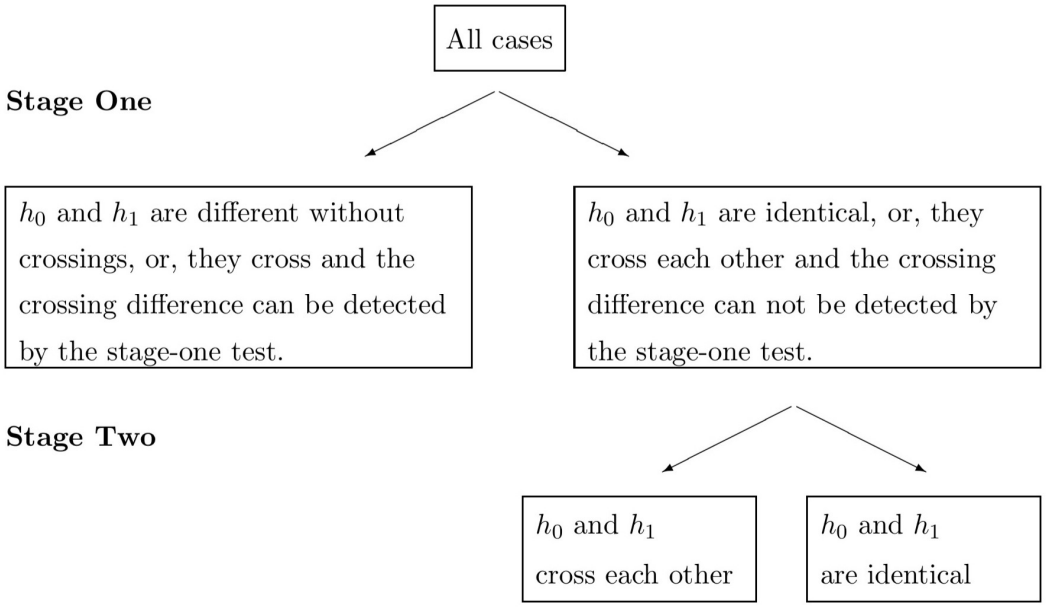


Figure 3.2: Diagram for demonstrating the Two-Stage testing procedure for comparing two hazard rates. h_0 and h_1 represent the hazard rate functions of survival times of subjects in the two groups that will be compared (Qiu and Sheng, 2008).

The Cox proportional hazards (PH) model is the most commonly used multivariate approach for analysing survival time data in medical research. Cox PH regression is a survival analysis regression model, which describes the relation between the event incidence, as expressed by the hazard function and a set of covariates, and does not directly model survival probabilities or survival times. A hazard rate (or failure rate) is the rate of occurrence of the event during a given time interval. A hazard ratio above 1 indicates a covariate that is positively associated with the event probability, and thus negatively associated with the length of survival (Bradburn *et al.*, 2003a; Schober and Vetter, 2018).

As in the conventional linear regression models, survival regression models allow for the quantification of the effect on survival of a set of predictors, the interaction of two predictors, or the effect of a new predictor above and beyond other covariates (Clark *et al.*, 2003).

However, the Cox proportional hazards model makes several assumptions. Thus, it is important to assess whether a fitted Cox regression model adequately describes the data. A survival model is adequate if it represents the survival patterns in the data to an acceptable degree. This aspect of a model is known as goodness of fit. Residuals are a useful method for

checking the fit of a statistical model. Essentially, they are the difference between an observed and a model-predicted quantity, with large or systematic differences between the two indicatives of a poor model. The proportional hazard assumption, that is, the hazards are proportional (and not overlapping) at all points in time, should be verified. For the Cox model, the (weighted) scaled Schoenfeld residuals test, the linear correlation test (Martingale residuals) and the time-dependent covariate test (Deviance residual) are the most powerful diagnostic tools for proportionality. The Deviance residual test is a symmetric transformation of the Martingale residuals test that assess influential observations. The first two test for an association between residuals and time (evidence of which indicates a bad fit), and the third tests whether the effect (coefficient) of a covariate changes with time (i.e. nonconstant hazard ratio). This latter method is appealing as it not only detects nonproportionality, but allows it to be modelled validly (Bradburn *et al.*, 2003b).

In this study, Kaplan-Meier plots and multivariate Cox proportional hazard models were used to examine the association between allelic expression and survival and were calculated using Survival package from R (Therneau, 2015; Therneau and Grambsch, 2000). Death due to all causes was used as endpoint for overall survival, and all subjects still alive were censored at the date of last contact. For Disease Specific Survival, only deaths due to breast cancer were used as endpoint, and deaths due to other causes and all subjects still alive at the date of last contact were censored. All Kaplan-Meier survival curves that did not cross were compared using the log-rank test using Survival package from R (Therneau, 2015; Therneau and Grambsch, 2000). When there was crossing of curves, the Two-Stage procedure was applied using the R package TSHRC (Qiu and Sheng, 2008), in which case the log-rank test was applied in stage one and, if not statistically significant, was followed by the stage two test. Since the TSHRC method only compares two curves, I did a pairwise comparison between MADE_wt, mut=wt and MADE_mut groups whenever necessary. P-values less than 0.05 were considered statistically significant.

For the multivariate analysis, Cox proportional hazard model was used to assess the effect of γ on overall and disease specific survivals. Hazard ratios (HRs) and 95% confidence intervals (CI) were estimated by fitting the Cox model while adjusting for age and tumour characteristics, such as size, Scarff-Bloom- Richardson histological grade, clinical stage and oestrogen receptor (ER), progesterone receptor (PR) and human epidermal growth factor receptor 2 (HER2) statuses. All p-values were two-sided, and p-values less than 0.05 were

considered statistically significant. The fit of the Cox model was verified with Schoenfeld residuals, Martingale residuals and Deviance residuals tests.

3.8 Graphical Representation

All plots were generated with ggplot2 package (Wickham, 2016), except for the survival curves, forest plots and the lollipops plots. The survival curves and forest plots were generated with Survminer R package that provides functions for facilitating visualization of survival analysis (Kassambara and Kosinski, 2018). For the mapping of mutations on linear proteins and its domains (lollipop plots), the MutationMapper online tool from cBioPortal was used (Cerami *et al.*, 2012; Gao *et al.*, 2013). The lollipop plots used for the introduction chapter were generated using the information from the 14 breast cancer datasets available in the cBioPortal. For the lollipop plots in the results chapter, information from the METABRIC and TCGA sets were used as input for the online tool.

CHAPTER 4
RESULTS

4 Results

4.1 METABRIC's set and TCGA's set mutations

In order to assess whether breast tumours harbouring somatic mutations displayed expression imbalance between the mutant and the wild-type alleles, mutant vs wild-type allelic expression analysis from tumour RNAseq data, from two independent sets of breast tumours, from the METABRIC and the TCGA projects was carried. Mutant allele ratios were determined for both DNA and RNA.

DNA allele ratios were used to determine differences in variant frequencies in DNA (**β** = $\log_2$[mutant allele DNA reads/wild-type allele DNA reads]), which was used to later normalise the RNA allele ratios. This allowed the analysis of two different levels of allelic expression read outs: total allelic expression imbalance (**α** = $\log_2$[mutant allele RNA reads/wild-type allele RNA reads]), and allelic expression imbalance due to cis-regulation alone (**γ**, equivalent to α normalised for different allele frequencies in DNA, i.e. [α]-[β]). In this way, α ratio reports on the general allelic expression imbalance, which can be generated by different mechanisms including copy-number aberrations and cis-regulatory variation, whilst γ ratio reports solely on allelic ratios due to cis-regulatory variation. Furthermore, I categorised the samples by γ value, into those with mutant allele differential expression (or MADE, $|\gamma| \geq 0.58$) and those without (equimolar expression of both alleles, mut=wt), which was required for the survival analysis carried later on.

In the METABRIC set, there were 255 somatic mutations in 25 genes. The two genes with more mutations where *PIK3CA* and *TP53*, with 115 and 69 tumours harbouring mutations on those genes, respectively. Only four genes where mutated in more than ten tumour samples (*PIK3CA*, *TP53*, *AKT1* and *GATA3*), and they all showed allelic imbalance. There were 14 genes mutated in more than two breast cancer samples, with only *ERBB3* not showing allelic imbalance. When assessing somatic mutations allelic imbalance in TCGA dataset, the two most mutated genes were, again, *PIK3CA* (n= 213) and *TP53* (n= 150). All the 116 genes mutated in more than ten breast cancer samples showed allelic imbalance (Figure 4.1 and Annex A).

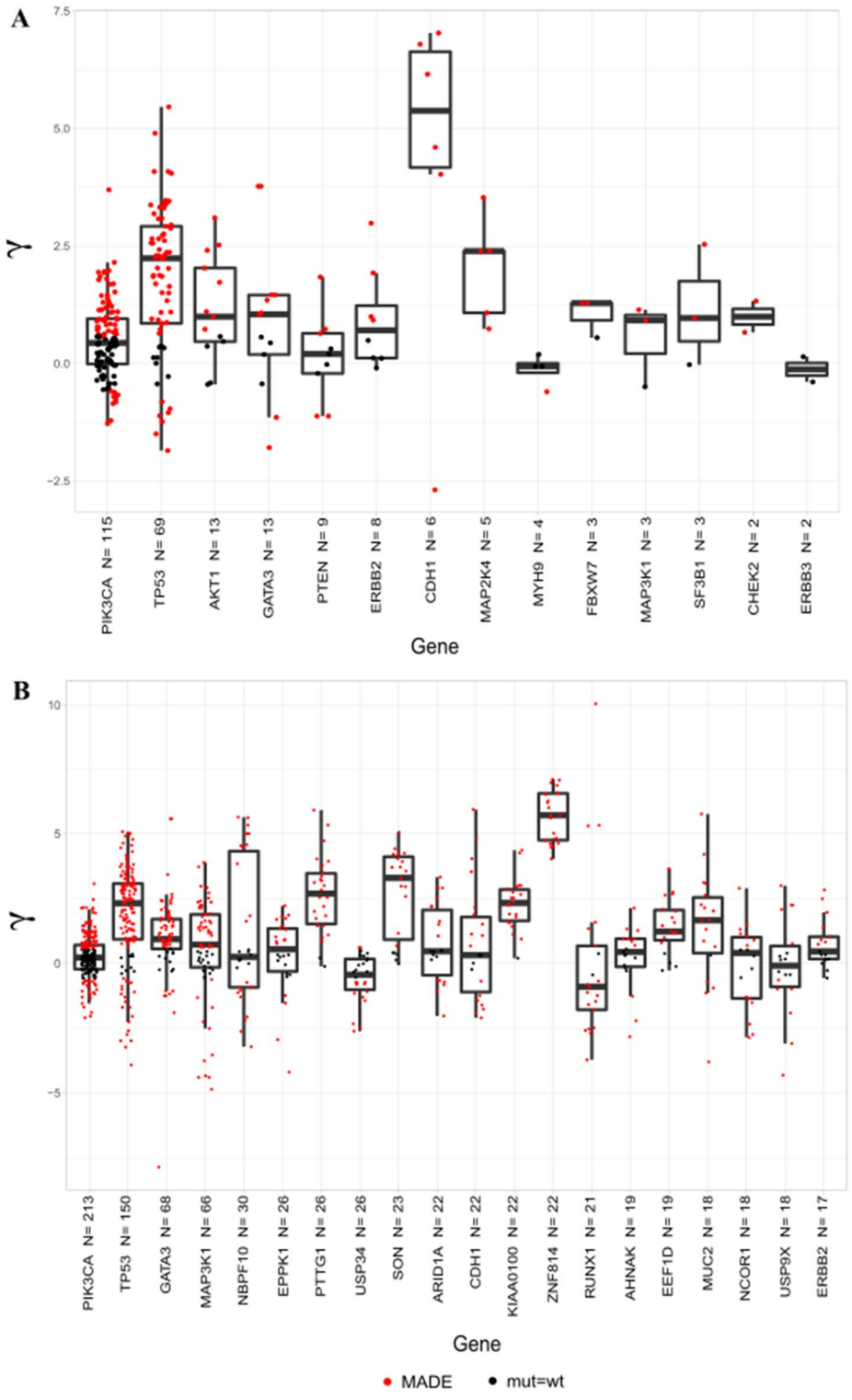


Figure 4.1: Somatic Mutations Allelic Imbalance. (A) γ ratios of all somatic mutated genes with more than one sample from METABRIC set. (B) γ ratios of the top 20 genes with more samples with somatic mutations from TCGA set. Each dot is a sample, and red dots correspond to samples with Mutant Allelic Differential Expression (MADE), while black dots are samples in which there is no allelic imbalance.

Since *PIK3CA* and *TP53* were the most mutated genes in both datasets, and our group has previously shown that they both have differential allelic expression in normal breast tissue, I set to analyse the association of *PIK3CA's* and *TP53's* somatic mutation allelic imbalance with breast cancer clinicopathological characteristics and patient outcome.

4.2 Mutant allele differential expression analysis of *PIK3CA* mutations in breast cancer

4.2.1 Protein Consequences of Mutations

First, I set to assess the impact of the *PIK3CA's* somatic mutations present in both METABRIC and TCGA sets. For that, I manually checked the OncoKB classification (Chakravarty *et al.*, 2017) of the amino acid alteration and then analysed if there was a difference in the distribution of the type of mutations' consequence between different MADE categories. I found that the majority of mutations were oncogenic and there was no statistical difference in the distribution of type of consequence between MADE categories (Figure 4.2 and Annex B). When looking at the distribution of the amino acid alterations on the protein, I observed that the majority of mutations occurred on the hotspots previously described (Bhat-Nakshatri *et al.*, 2016).

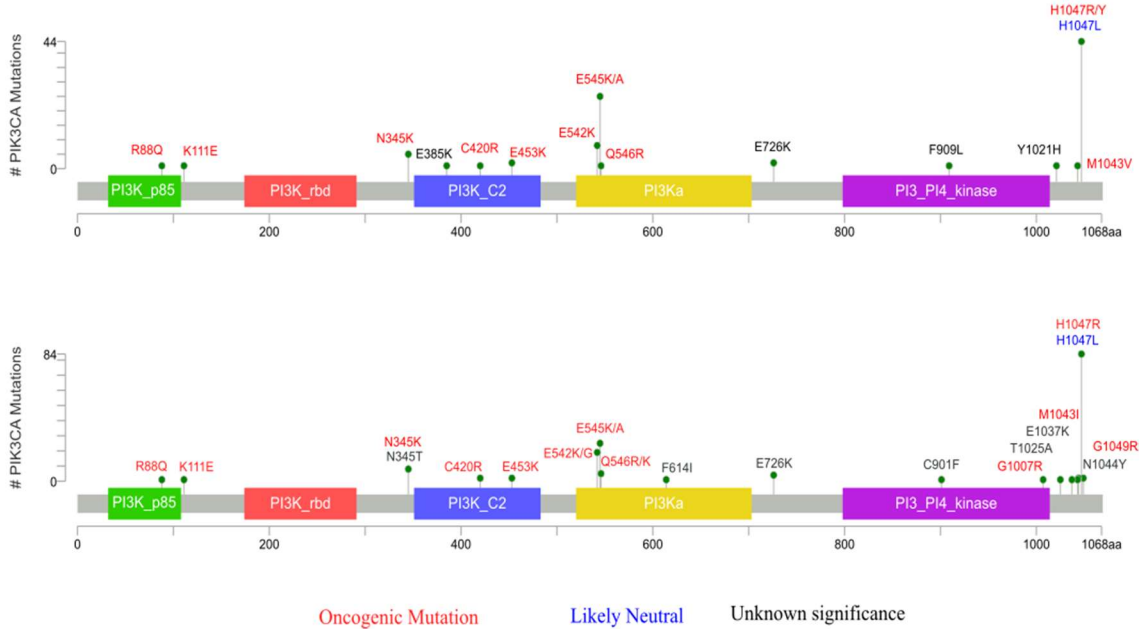


Figure 4.2: Amino-acid alterations in p110α protein, encoded by the *PIK3CA* gene. The distribution of missense mutations in amino-acids in p110α in METABRIC set (A) and in TCGA set (B). Green lollipops represent the missense mutations, with the height of the lollipop representing the total number of each mutation. Amino acid changes written in red represent the oncogenic mutations, in blue likely neutral mutations and in black mutations with unknown significance.

4.2.2 Differential Allelic Expression of *PIK3CA*'s Somatic Mutations in Breast Cancer

Next, I set out to assess whether *PIK3CA* somatic mutations would have their penetrance modified by imbalances in allelic expression, and if so how significant the contribution from cis-regulation was. My premise was that gain-of-function mutations occurring on preferentially expressed alleles would have a different impact, than those occurring in lowly expressed alleles. To this end, mutant vs wild-type allelic expression analysis was carried for breast tumour samples carrying somatic *PIK3CA* missense mutations, on two independent sets of data, the METABRIC (METABRIC set, n=94) and the TCGA (TCGA set, n=161) projects. The summary description of the two sets can be found in Annex C.

When looking at the distribution of α, β and γ ratios of *PIK3CA's* somatic mutations in breast cancer tumours in both datasets, I observed that the average γ ratio (0.5809 and 0.2646, for METABRIC and TCGA sets respectively) was significantly higher than the average of the other two ratios (pairwise Mann-Whitney U test p-value < 2e-16), meaning that the mutant allele is preferentially expressed when taking into account cis-regulation alone (Figure 4.3 and Annex D).

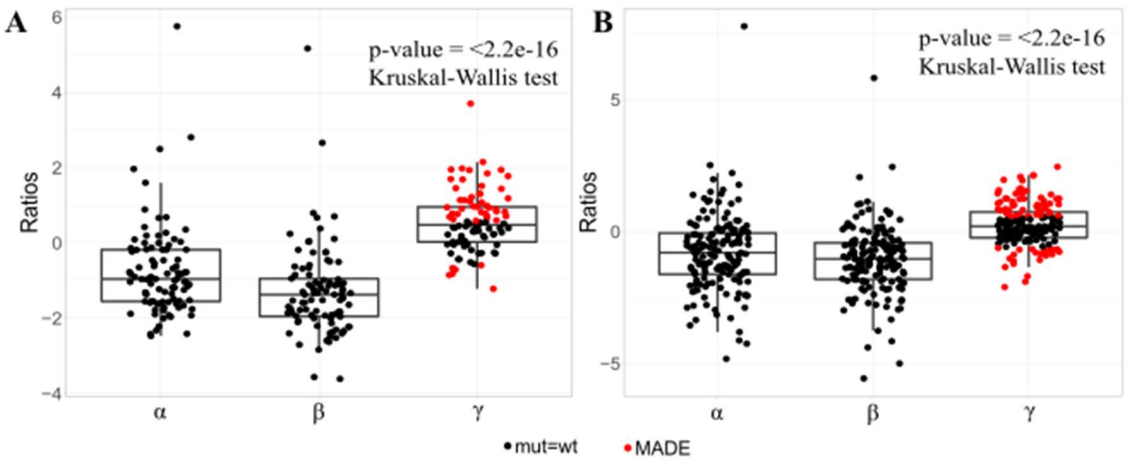


Figure 4.3: Distribution of *PIK3CA's* α, β and γ ratios in breast tumours in (A) METABRIC set and (B) TCGA set. Average γ was significantly higher than α and β (p-value corresponds to Kruskal-Wallis test).

There was a high percentage of tumours displaying total mutant vs wild-type allelic expression imbalances (α ratios) greater than 1.5-fold, both in the METABRIC set (75.53%) and in the TCGA set (66.46%). The same was true for the normalized expression allelic mutant/wild-type ratios (γ ratios) which showed a small decrease in the percentages when

compared to the non-normalized ratios (50% and 46.59%, for METABRIC and TCGA sets, respectively; Table 4.1).

When looking at the distribution of expression imbalance values in both sets, tumours with preferential expression of the mutant allele displayed higher fold-change differences than tumours with higher expression of the wild-type allele. This was observed for total expression, but also for normalized ratios corrected for variant frequencies in DNA (Table 4.1).

Table 4.1: Summary of allelic imbalance of *PIK3CA's* somatic mutations in breast tumours

	METABRIC set (n=94)		TCGA set (n=161)	
	α	γ	α	γ
Mutation allelic imbalance				
absolute ratio ≥ 0.58 (%)	75.53	50	66.46	46.59
Mut Allele > WT Allele (%)	8.51	43.62	11.18	32.3
WT Allele > Mut Allele (%)	67.02	6.38	55.28	14.29
Maximum Fold Difference				
Mutant Allele	54	13.01	220	5.49
Wild-type Allele	5.53	2.32	28.33	4.29

4.2.3 Positive selection for cis-regulated preferentially expressed mutated alleles

Changes in copy number in tumours have been previously associated with changes of gene expression in cis (Soh *et al.*, 2009; Curtis *et al.*, 2012; Hartman *et al.*, 2012; Krasinskas *et al.*, 2013; Bielski *et al.*, 2018), but it has never been analysed whether differential expression of a mutation could also be due to cis-regulation. To dissect the contribution of both cis-regulation and chromosomal copy number level to these differences in mutant allelic expression, I investigated whether there was positive, neutral or negative selection for the mutant allele imbalance, by comparing γ and β values for each tumour, in both sets. I found that, in both cohorts, the majority of samples (71.28% and 55.28% for the METABRIC and TCGA sets, respectively) had positive γ and negative β values (blue dots on Figure 4.4), indicating positive selection for cis-regulation, and not for allele count (Figure 4.4). These samples showed preferential expression of the mutant allele, albeit the larger proportion of the wild-type allele at the DNA level.

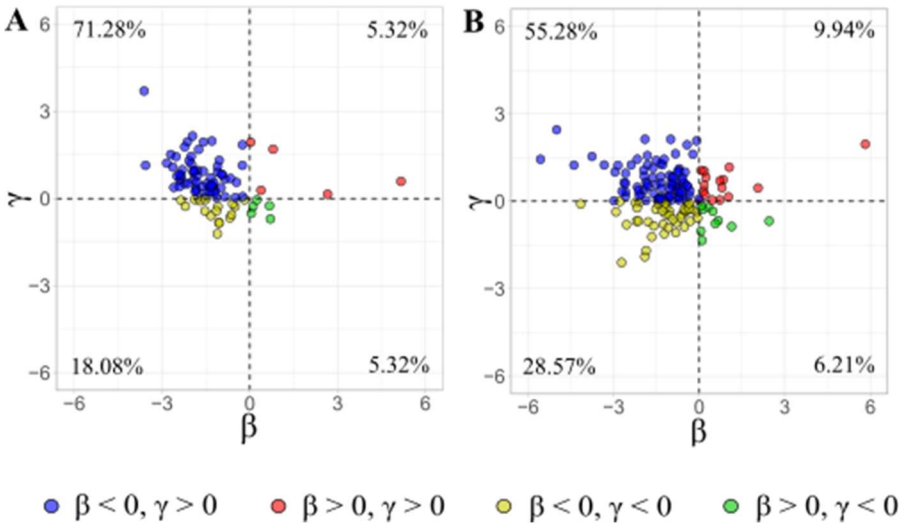


Figure 4.4: Comparison of matched *PIK3CA's* β and γ values showing predominance of tumours showing preferential allelic expression of the mutated allele, albeit higher number of wild-type allele copies.

Surprisingly, only a small fraction of samples displayed preferential allelic expression and higher allele copy number of the mutation (5.32% and 9.94% for the METABRIC and TCGA sets, respectively; red dots on Figure 4.4). To evaluate how much of these results might be confounded by different levels of cellularity or tumour purity, I introduced this information in the analysis and found no association between these parameters and the β or γ values (Figure 4.5). Altogether, these results indicate that cis-regulation of expression of the mutant allele is positively selected for in tumours, and that it is predominant to allelic copy number variation in determining preferential expression of mutations in tumours. Since I did not have information on tumour purity for TCGA set, I could not perform the same analysis for such set.

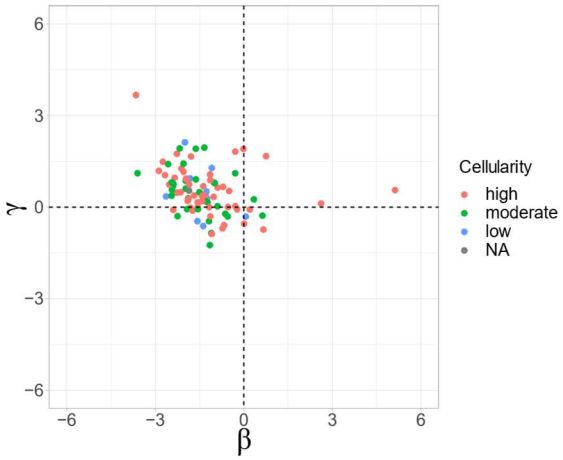


Figure 4.5: Comparison of *PIK3CA's* matched β and γ values dots coloured by cellularity levels showing that there is no pattern of association between ratios and cellularity

4.2.4 MADE defines an aggressive subset of *PIK3CA* mutated tumours

To investigate the impact of differential cis-regulation of *PIK3CA*'s mutations on clinical outcome, I performed univariate survival analysis comparing the γ ratios with respect to overall survival and also to disease specific survival. As Kaplan-Meier analysis only accommodates categorical variables, I used the MADE categorisation. I found that the MADE group had poorer overall survival rates than the mut=wt group (p-value = 0.03), with the median survival for the MADE group at approximately 8 years and the mut=wt group at 18.5 years (Figure 4.6 and Annex E), in the METABRIC set. When analysing disease specific survival, the MADE group also had poorer survival rates when compared with the mut=wt (p-value= 0.012), albeit both groups not reaching the median survival (Figure 4.6 and Annex E). When further dividing MADE group into tumours that expressed more of the mutated allele (MADE_mut) and tumours that expressed more of the wild-type allele (MADE_wt) and applying survival analysis, I observed that the MADE_wt group was associated with poorer overall and disease specific survivals (Figure 4.6). When performing pairwise comparison of the curves, for overall survival analysis, the only significant difference was between MADE_wt and mut=wt groups (p-value = 0.0005). However, for disease specific survival there was a significant difference between MADE_wt and mut=wt groups (p-value = 0.002), and also between MADE_mut and mut=wt groups (p-value = 0.03, Table 4.2).

Subsequently, in a multivariate analysis of overall survival and disease specific survival, in which I fit a Cox model while adjusting for age and tumour characteristics, such as size, Scarff-Bloom- Richardson histological grade, clinical stage and ER, PR and HER2 statuses, γ ratios were not significantly associated with neither types of survival, when used as a continuous variable nor when used MADE categorization (Annex I). However, some of the variables that usually are independent prognostic factors, such as PR and HER2 statuses were also not associated with hazard of breast cancer.

When performing diagnostic for the Cox regression model, the proportional hazard assumption was not violated, there was no influential observation and there was a non-linear relation between the log hazard and the variables (Annex I). Together these results show that the Cox model is appropriate for the data.

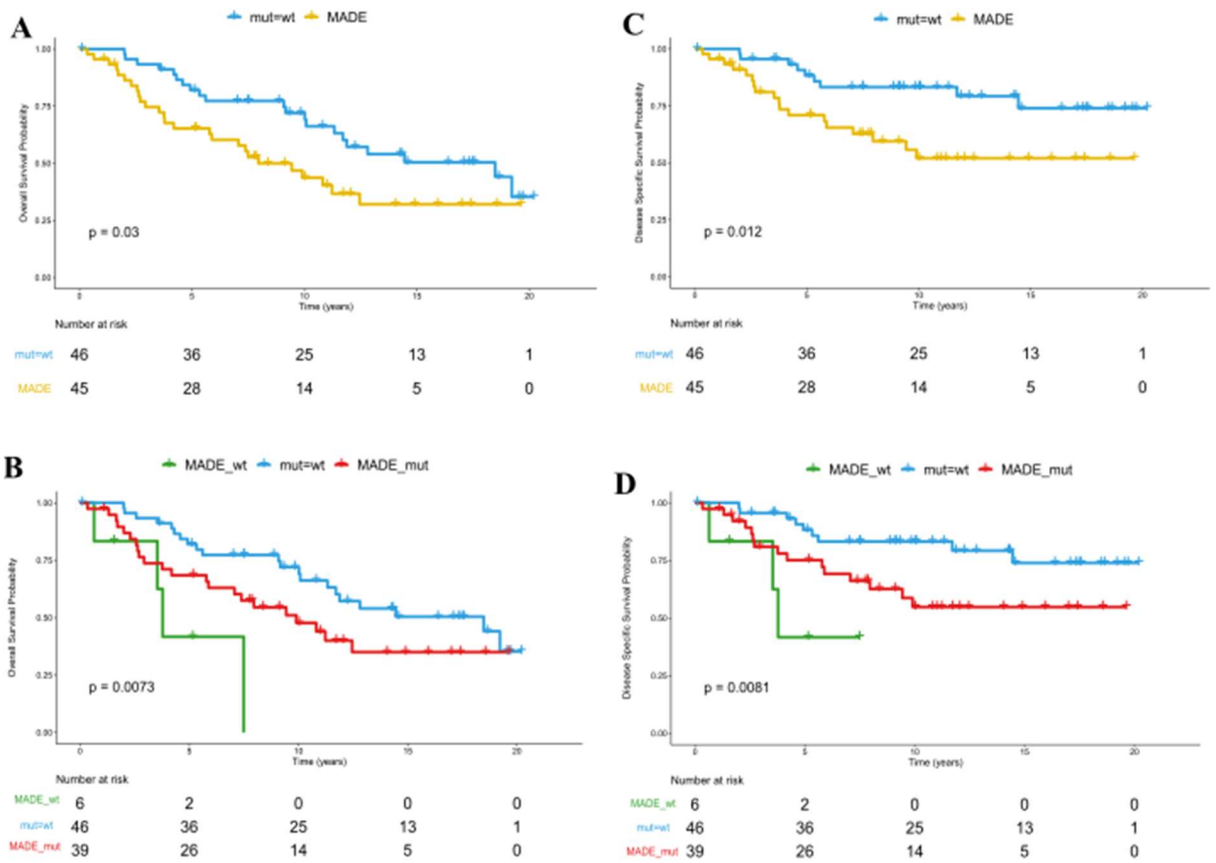


Figure 4.6: Kaplan-Meier analysis of overall survival and disease specific survival of *PIK3CA's* MADE in the METABRIC set. (A) Overall Survival of METABRIC set divided in MADE and mut=wt groups. (B) Overall Survival with MADE group further divided in tumours that express more wild-type allele (MADE_wt) and tumours that express more mutated allele (MADE_mut). (C) Disease Specific Survival of METABRIC set divided in MADE and mut=wt groups. (D) Disease Specific Survival with MADE group further divided in tumours that express more wild-type allele (MADE_wt) and tumours that express more mutated allele (MADE_mut).

Table 4.2: P-values of pairwise comparison of overall and disease specific survivals of *PIK3CA's* MADE in METABRIC set (p-values calculated by log-rank test if curves do not cross and by Two-Stage procedure if curves cross. OS: Overall Survival. DSS: Disease specific survival).

	OS	DSS
MADE_wt : mut=wt	0.0005	0.00228
MADE_wt : MADE_mut	0.6534	0.9793
mut=wt : MADE_mut	0.1821	0.0305

In the TCGA set, there was no significant difference in overall or in disease specific survivals between the MADE and mut=wt group, nor with further subdivision of MADE group, even when applying pairwise test between groups (Figure 4.7 and Table 4.3, Annex G).

Kaplan-Meier analysis was also performed for α ratios in both datasets. In order to achieve that, first, the α ratios where divided into α_DAE and α_noDAE groups. There was no statistically significant difference between survival of the two groups for the METABRIC set when considering overall survival (p-value = 0.71), nor when taking into account disease specific survival (p-value = 0.76, Figure 4.8). For the TCGA set, there was no difference in the survival of the two groups for overall survival, but the disease specific survival showed the α_noDAE group with poorer survival, when compared with the α_DAE group (p-value = 0.011), with α_noDAE group's median survival of 9.2 years and α_DAE group not reaching median survival and having more than 50% of patients surviving past the end of observation period (Figure 4.9 and Annex G).

For the next analysis I further divided the α_DAE group in two, α_DAEwt (preferential expression of wild-type allele) and α_DAEmut (preferential expression of mutated allele). For the METABRIC set neither overall nor disease specific survival analysis showed significant difference between groups (p-value = 0.41 and 0.14, respectively, Figure 4.8), and when applying pairwise comparison, there was also no significant difference (Table 4.4). In the TCGA set, there was no difference between groups when assessing overall survival, but the disease specific survival was significant, with the α_noDAE group being the one with worst outcome (p-value = 0.034, Figure 4.9). When comparing the curves in pairs, I observed that the only significant difference was between α_DAEwt and α_noDAE groups (P-value = 0.009, Table 4.5), with α_noDAE group's median survival of 9.2 years and the α_DAEwt group not reaching median survival point by the end of the observation period (Annex G).

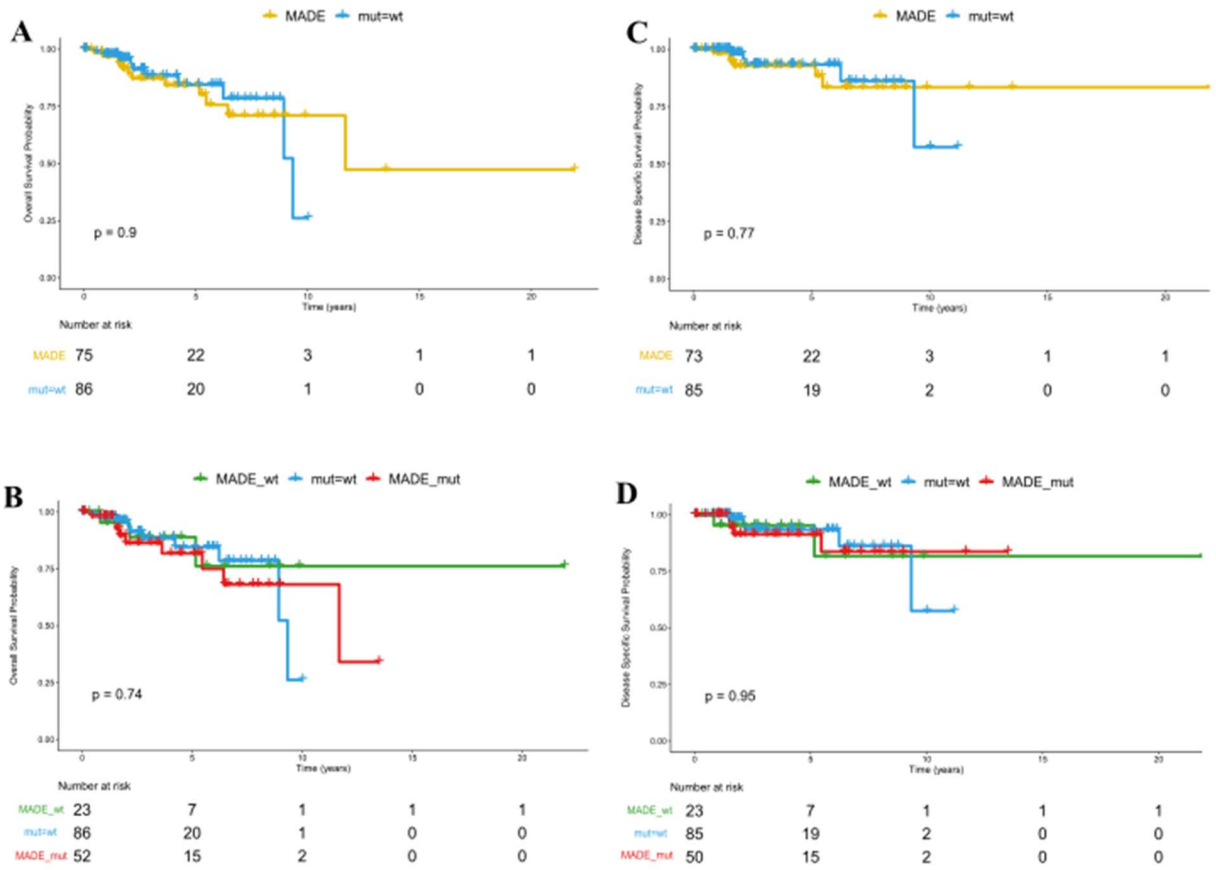


Figure 4.7: Kaplan-Meier analysis of overall survival and disease specific survival of *PIK3CA's* MADE in the TCGA data set. (A) Overall Survival of TCGA set divided in MADE and mut=wt groups. (B) Overall Survival with MADE group further divided in tumours that express more wild-type allele (MADE_wt) and tumours that express more mutated allele (MADE_mut). (C) Disease Specific Survival of TCGA set divided in MADE and mut=wt groups. (D) Disease Specific Survival with MADE group further divided in tumours that express more wild-type allele (MADE_wt) and tumours that express more mutated allele (MADE_mut).

Table 4.3: P-values of pairwise comparison of overall and disease specific survivals of *PIK3CA's* MADE in TCGA set (p-values calculated by log-rank test if curves do not cross and by Two-Stage procedure if curves cross).

	OS	DSS
MADE_wt : mut=wt	0.5143	0.842
MADE_wt : MADE_mut	0.9076	0.994
mut=wt : MADE_mut	0.4854	0.4425

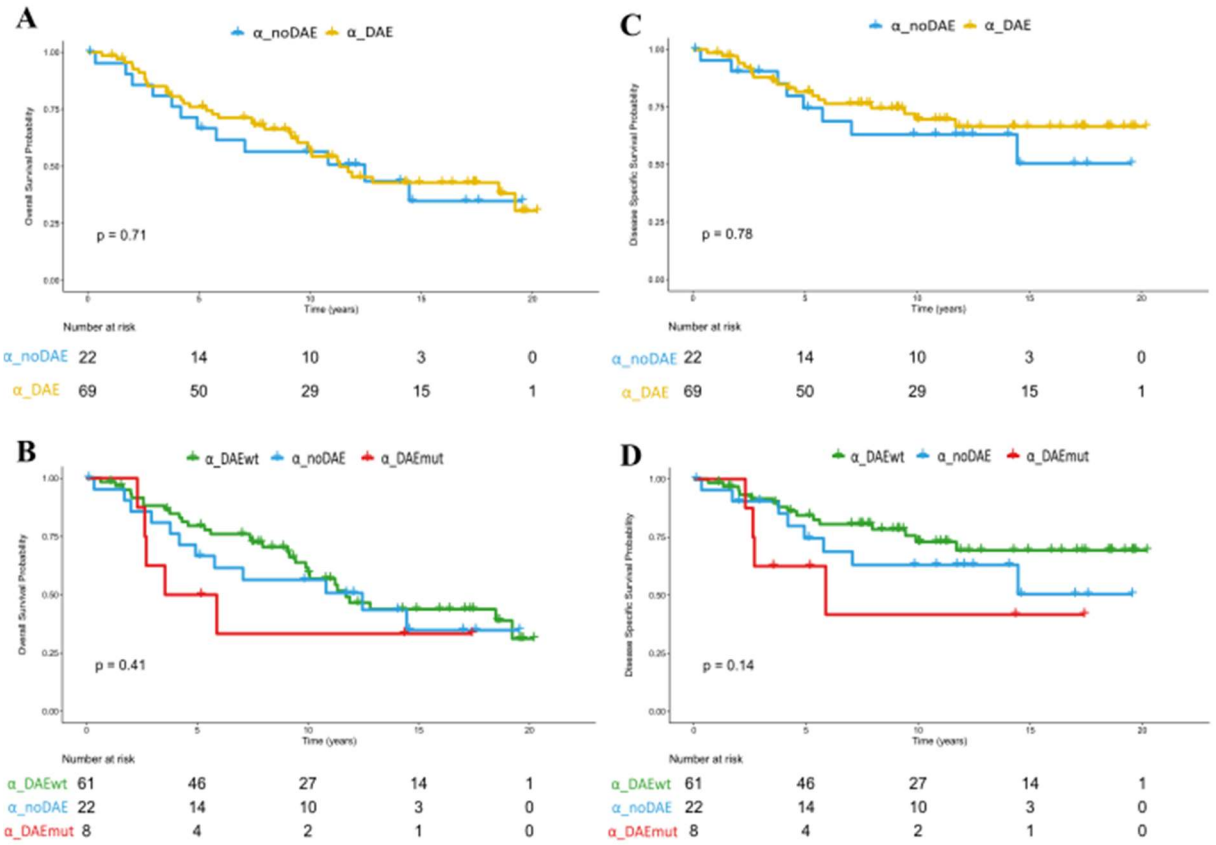


Figure 4.8: Kaplan-Meier analysis of overall survival and disease specific survival of *PIK3CA's* α_DAE in the METABRIC data set. (A) Overall Survival of METABRIC set divided in α_DAE and α_noDAE groups. (B) Overall Survival with α_DAE group further divided in tumours that express more wild-type allele (α_DAEwt) and tumours that express more mutated allele (α_DAEmut). (C) Disease Specific Survival of METABRIC set divided in α_DAE and α_noDAE groups. (D) Disease Specific Survival with α_DAE group further divided in tumours that express more wild-type allele (α_DAEwt) and tumours that express more mutated allele (α_DAEmut).

Table 4.4: P-values of pairwise comparison of overall and disease specific survivals of *PIK3CA's* α_DAE in METABRIC set (p-values calculated by log-rank test if curves do not cross and by Two-Stage procedure if curves cross).

	OS	DSS
α_DAEwt : α_noDAE	0.5594	0.7668
α_DAEwt : α_DAEmut	0.2161	0.8139
α_noDAE : α_DAEmut	0.8866	0.8347

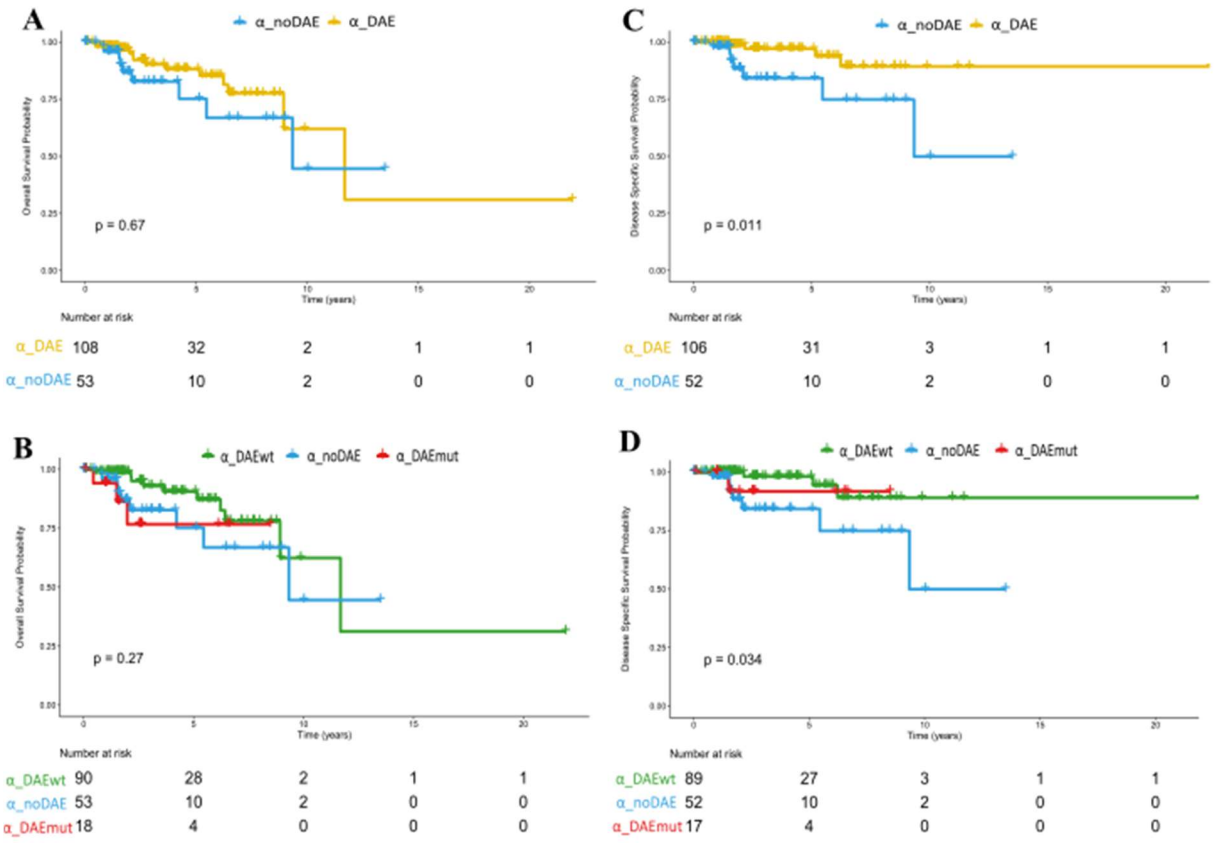


Figure 4.9: Kaplan-Meier analysis of overall survival and disease specific survival of *PIK3CA's* α_DAE in the TCGA data set. (A) Overall Survival of TCGA set divided in α_DAE and α_noDAE groups. (B) Overall Survival with α_DAE group further divided in tumours that express more wild-type allele (α_DAEwt) and tumours that express more mutated allele (α_DAEmut). (C) Disease Specific Survival of TCGA set divided in α_DAE and α_noDAE groups. (D) Disease Specific Survival with α_DAE group further divided in tumours that express more wild-type allele (α_DAEwt) and tumours that express more mutated allele (α_DAEmut).

Table 4.5: P-values of pairwise comparison of overall and disease specific survivals of *PIK3CA's* α_DAE in TCGA set (p-values calculated by log-rank test if curves do not cross and by Two-Stage procedure if curves cross).

	OS	DSS
α_DAEwt : α_noDAE	0.1865	0.0089
α_DAEwt : α_DAEmut	0.1101	0.7569
α_noDAE : α_DAEmut	0.484	0.97337

4.2.5 *PIK3CA* MADE correlates with clinicopathological variables

Next, the clinical impact of *PIK3CA*'s allelic imbalance on tumour biology was explored by correlating the mutant/wild-type expression ratios with breast cancer clinicopathological characteristics, such as the hormonal receptor and HER2 statuses, PAM50 and Integrative Clusters classifications, tumour size, stage, histological grade, lymph node

status and patients age, for the METABRIC set. To accomplish this, the allelic expression ratios were used as a continuous variable and applied a bivariate analysis, whose main results are shown in Figure 4.10, Figure 4.11 and Annex J.

The mean of the α ratios in ER negative tumours was significantly higher than in ER positive tumours (p-value=0.025), meaning higher expression of mutant allele in ER negative tumours. The same was found for PR negative tumours, which showed, on average, a higher expression of the mutant allele, when compared with PR positive tumours (p-value=0.007). To evaluate the effect of cis-regulation in this association, the same analysis was carried out for γ ratios and found the same association for PR expression (p-value=0.039), but there was no significant association for ER expression (p-value=0.129).

A significant association between the allelic expression ratios and HER2 status was also found, but in the opposite direction, as the mean α and γ ratios were higher in HER2 positive when compared with HER2 negative tumours (p-value=0.018 and p-value=0.025, respectively).

The analysis of the TCGA set, confirmed the significant association between the PR status and expression with both α and γ ratios (p-values=0.045 and p-value=0.002, respectively). It also showed an association between ER status and γ (p-value=0.002), but not with α ratios. There was no significant association for HER2 status. (Annex J, Figure 4.10 and Figure 4.11)

There was no association between the α or the γ ratios and the other clinical variables, such as PAM50 and Integrative Cluster classifications, age, stage and lymph node status (Annex J).

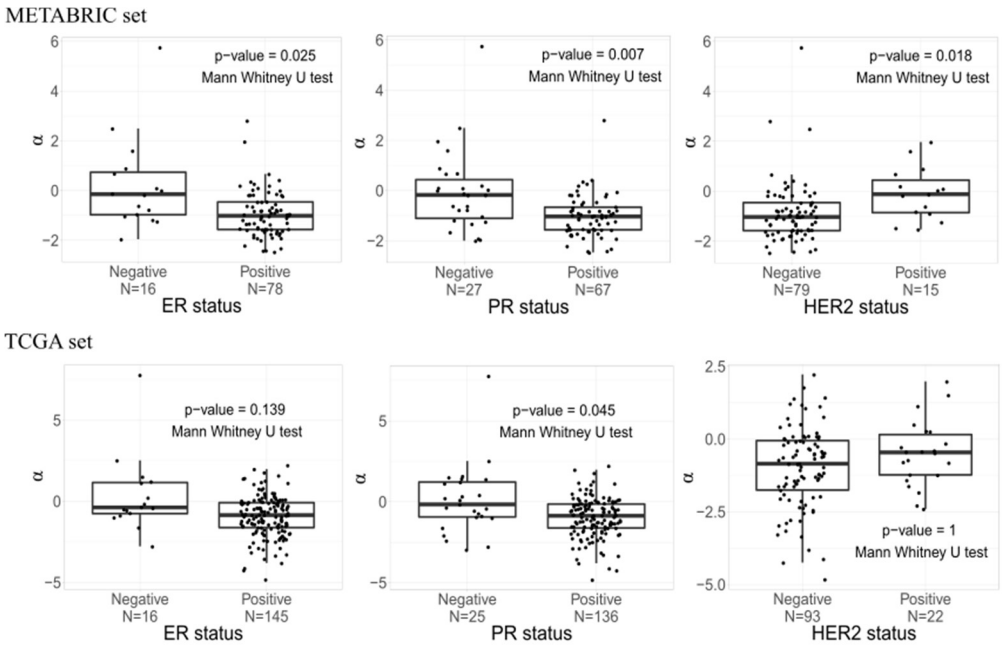


Figure 4.10: Association of *PIK3CA's* α ratios and receptors statuses. Upper Line: Boxplots showing association between α ratios and ER, PR and HER2 statuses, respectively for the METABRIC set. Bottom line: Boxplots showing association between α ratios and ER, PR and HER2 statuses, respectively for the TCGA set. Each dot represents a sample.

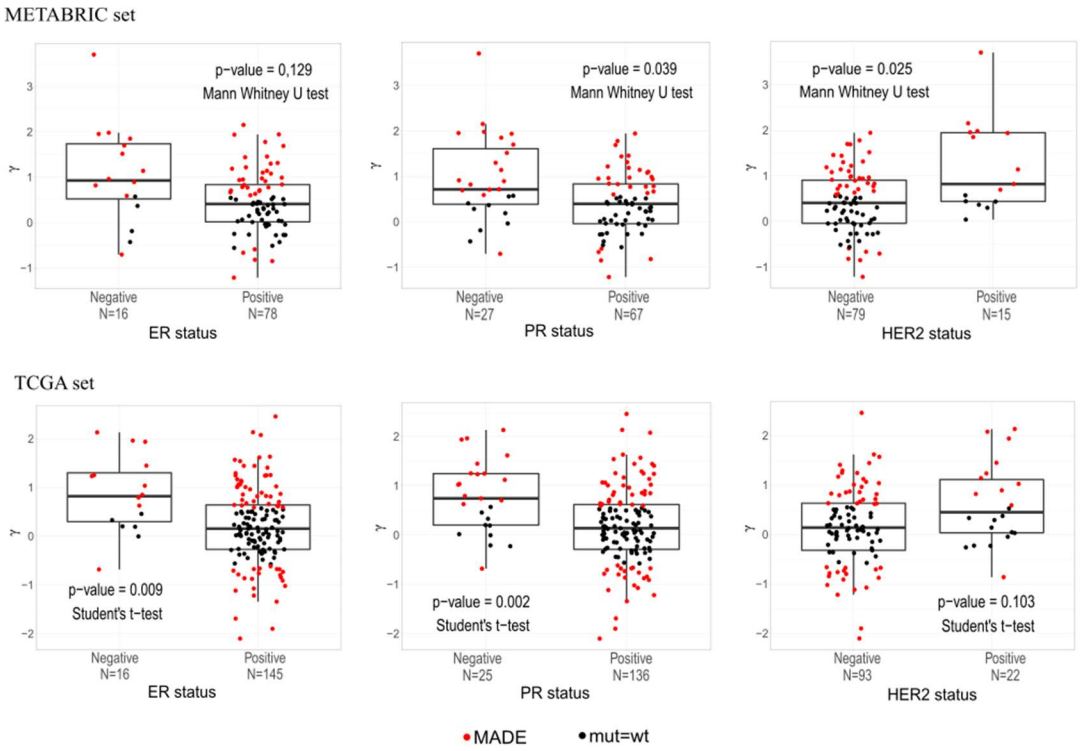


Figure 4.11: Association of *PIK3CA's* γ ratios and receptors statuses. Upper line: Boxplots showing association between γ ratios and ER, PR and HER2 statuses for the METABRIC set. Bottom line: Boxplots showing association between γ ratios and ER, PR and HER2 statuses for the TCGA set. Each dot is a sample. Red dots correspond to tumours with MADE, while black dots represent samples with no allelic imbalance.

To assess whether the distribution of patient age and tumours clinicopathological characteristics were different between MADE groups, chi-square tests were performed for the METABRIC and for the TCGA sets. There was no significant difference in distribution of tumour's grade, size and stage, Integrative Clusters and PAM50 classifications, number of positive lymph nodes, age, ER status or HER2 status between different MADE groups for neither datasets. Distribution of PR status was only significantly different in the TCGA set (p-value = 0.02), with the MADE_mut group being comprised of more PR negative samples than mut=wt group (p-value = 0.021, Annex K).

Since, I observed associations between ER, PR and HER2 statuses and γ ratios, I set to explore the effect of γ ratios in survival, according to these receptor statuses. First, the datasets were divided into receptor positive and receptor negative groups, then Kaplan-Meier overall and disease specific survival analyses were applied. When assessing the disease specific survival, I only observed a statistically significant difference in the HER2 negative group of the METABRIC set, with the MADE group presenting a poorer survival when compared to mut=wt group (p-value = 0.027; Annex F). When further dividing MADE group into MADE_wt and MADE_mut and analysing the disease specific survival, I observed statistically significant differences in ER positive, PR positive and HER2 negative groups, with MADE_wt having the worst outcome (p-values = 0.039, 0.026 and 0.0093, respectively; Figure 4.12). When looking at overall survival, there were statistically significant differences in survival in ER positive, PR positive and HER2 negative groups, when dividing these groups into MADE_wt, MADE_mut and mut=wt groups. Again MADE_wt group had poorer survival (p-value = 0.004, 0.0013 and 0.0066, for ER positive, PR positive and HER2 negative respectively, Annex F).

When comparing the curves in a pairwise way, I also found significant difference in PR negative group, for overall survival with the significant difference being between MADE_wt and MADE_mut (p-value = 0.031), and for disease specific survival between MADE_wt and mut=wt (p-value = 0.026), with MADE_wt presenting poorer survival. In the ER positive group, significant overall and disease-specific differences were detected between MADE_wt and mut=wt (p-value = 0.001 and 0.019, respectively) and between MADE_wt and MADE_mut (p-value = 0.007 and 0.046, respectively). In the PR positive group, a significant overall and disease-specific survival difference was also found between MADE_wt and mut=wt (p-value = 0.0005 and 0.016, respectively) and between MADE_wt and MADE_mut (p-value = 0.007 and 0.05, respectively). In the HER2 negative group, I found a significant difference for overall and disease specific survivals only between MADE_wt and mut=wt groups (p-value = 0.0007

and 0.002, respectively; Table 4.6 and Annex F). Both in ER positive, PR positive and HER2 negative groups mut=wt group presented better outcome and MADE_wt presented worse outcome.

For the TCGA set, there was no statistically significant difference in overall survival nor in disease specific survival when analysing MADE groups by receptor (Annexes N and O).

When assessing survival for the subgroups divided by receptors status, taking into account α ratios, there was no statistically significant difference in overall nor disease specific survival in METABRIC set. The exception was a significant difference in overall survival in PR positive subgroup with α_DAEwt presenting shorter median survival when compared to α_DAEmut group (p -value = 0.025; Annexes E and F), but α_DAEmut group is comprised of only one subject that was censored (did not die during period of observation). For TCGA set, the only significant differences observed were in disease specific survival of ER positive and PR positive subsets, with α_DAE group presenting better survival than α_noDAE group (p-value = 0.039 and 0.018, respectively. Figure 4.13 and Annex H)

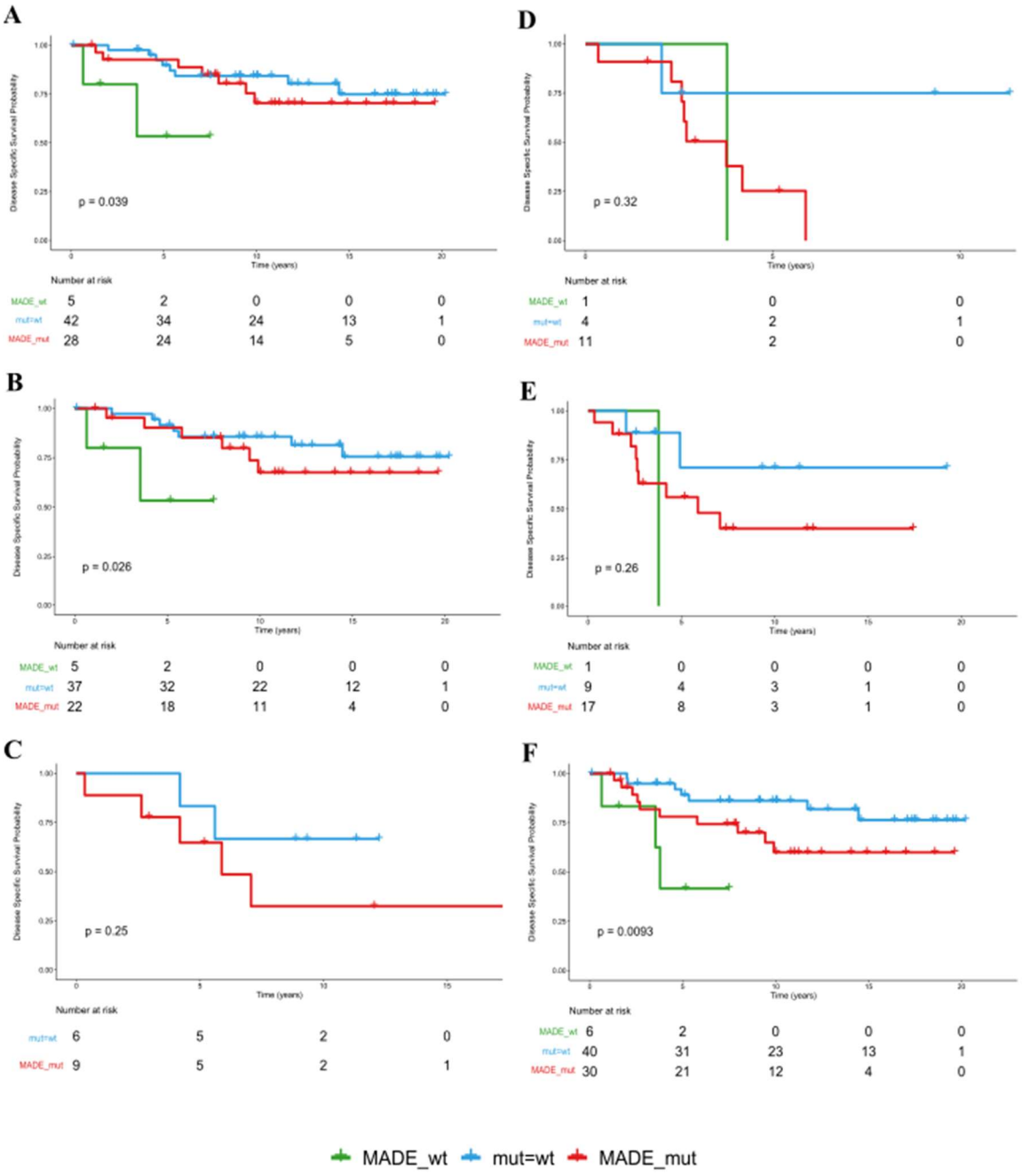


Figure 4.12: Kaplan-Meier analysis of disease specific survival of *PIK3CA's* MADE groups in METABRIC set according to ER, PR and HER2 statuses. Disease specific survival in (A) ER positive group, (B) PR positive group, (C) HER2 positive group, (D) ER negative group, (E) PR negative group and (F) HER2 negative group.

Table 4.6**:** P-values of pairwise comparison of disease specific survivals of *PIK3CA's* MADE in METABRIC set divided according to receptor status (p-values calculated by log-rank test if curves do not cross and by Two-Stage procedure if curves cross).

	ER positive	ER negative	PR positive	PR negative	HER2 positive	HER2 negative
MADE_wt : mut=wt	0.0193	0.101	0.0166	0.026		0.00225
MADE_wt : MADE_mut	0.046	0.2455	0.05	0.1395	0.25	0.7336
mut=wt : MADE_mut	0.5855	0.6857	0.6227	0.6672		0.078

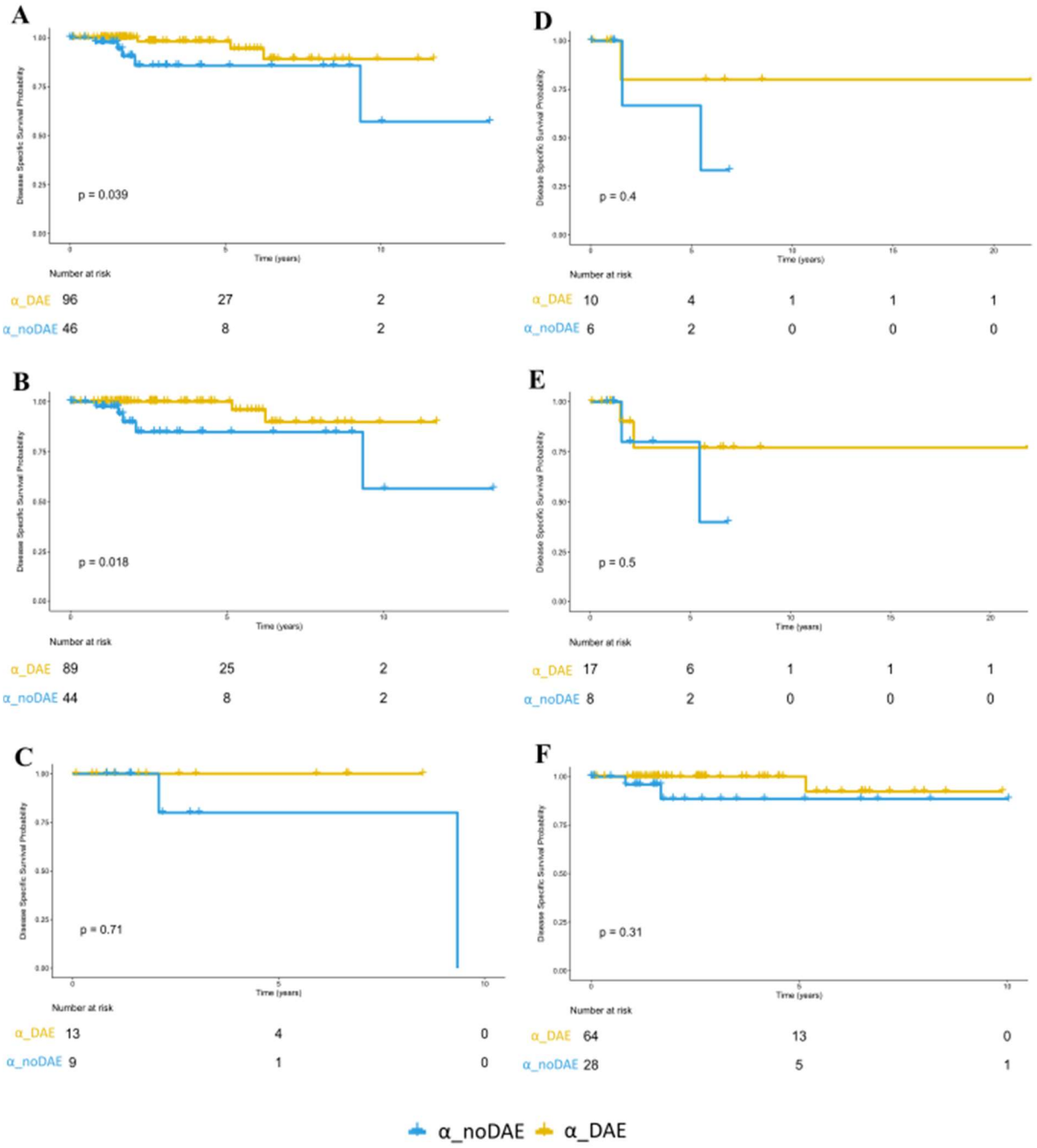


Figure 4.13: Kaplan-Meier analysis of disease specific survival of *PIK3CA's* α_DAE groups in TCGA set according to ER, PR and HER2 statuses. Disease specific survival in (A) ER positive group, (B) PR positive group, (C) HER2 positive group, (D) ER negative group, (E) PR negative group and (F) HER2 negative group.

4.2.6 *PIK3CA's* regulatory SNP

Our group had previously analysed normal breast tissue data and identified one SNP, rs2699887, as the only strong cis-regulatory candidate SNP for the observed *PIK3CA's* differential allelic expression. The group also detected that this SNP was associated with

differences in the total expression of *PIK3CA* (p-value= 0.011) in tumours from the METABRIC set (manuscript under preparation). rs2699887 locates at the first intron of *PIK3CA*, in a region rich in epigenetic marks and classified as an active promoter in breast mammary epithelial cells (vHMEC) and breast myoepithelial primary cells. The rs2699887 has two alleles: the C allele which is the major allele (the most commonly found allele in the population); and the T allele, which is the minor allele. There is an *in silico* prediction that the minor allele would lead to higher expression of *PIK3CA*. Thus, I set to analyse the possible association between rs2699887 genotype and patients' survival.

First, I set to analyse if there was a difference in the distribution of *PIK3CA's* MADE between the different genotypes (TT, CT, CC). I observed that there was no difference in the distribution of MADE groups between subtypes (chi-square test p-value = 0.21; Figure 4.14).

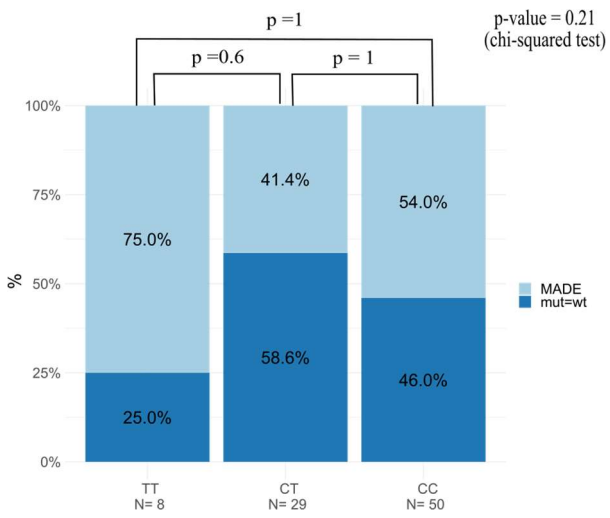


Figure 4.14: Distribution of *PIK3CA's* MADE between the different rs2699887 genotypes groups (TT, CT, CC). P-values correspond to chi-square test, and pairwise chi-square tests were adjusted with Bonferroni's correction.

Then, I performed overall and disease specific survival analysis for the three possible genotypes (TT, CT, CC). I did not observe any significant difference in overall or disease specific survival (Annex M).

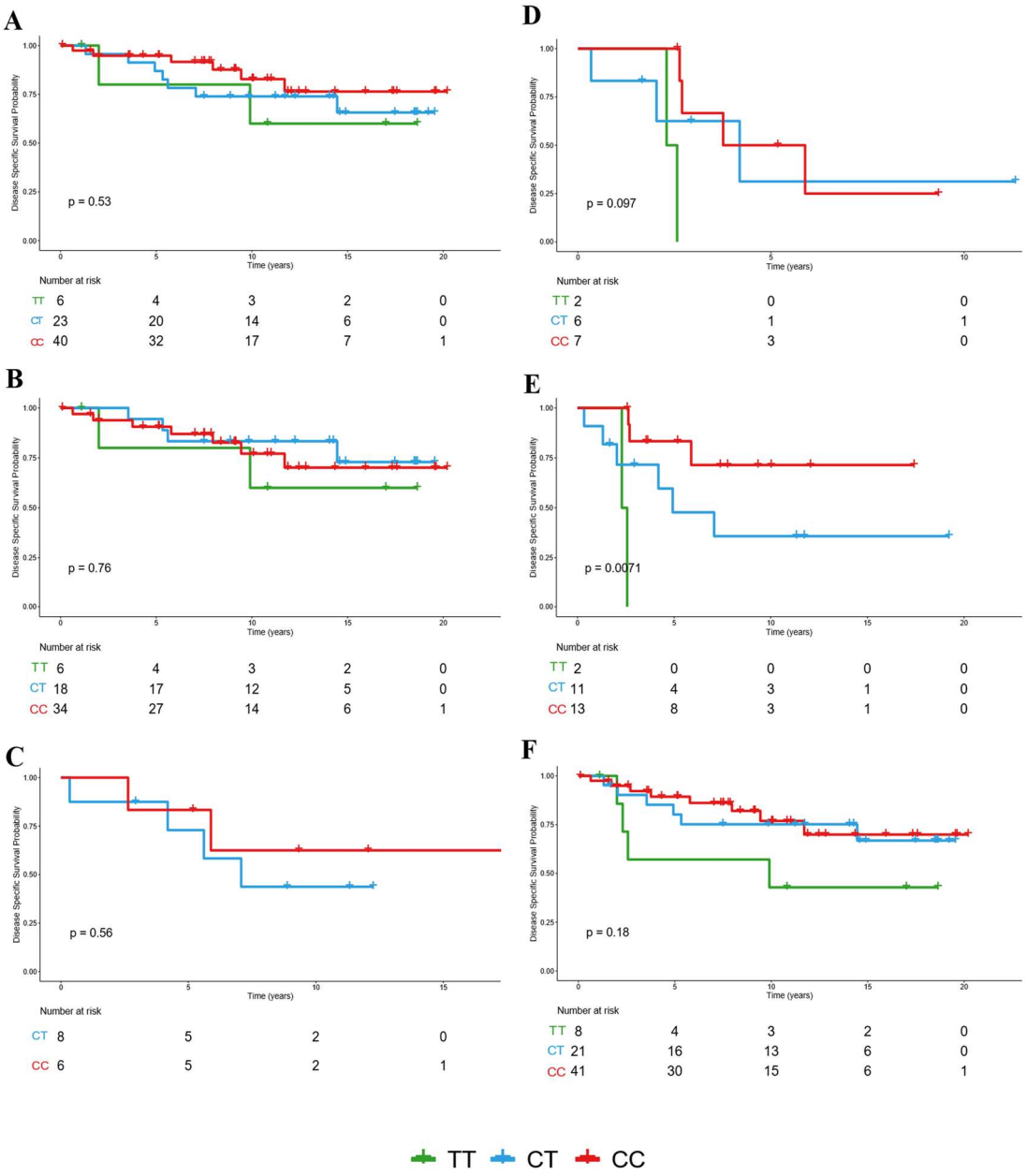


Figure 4.15: Kaplan-Meier analysis of disease specific survival of *PIK3CA's* regulatory SNP genotype TT, CT and CC groups in METABRIC set according to ER, PR and HER2 statuses. Disease specific survival in (A) ER positive group, (B) PR positive group, (C) HER2 positive group, (D) ER negative group, (E) PR negative group and (F) HER2 negative group.

Table 4.7: P-values of pairwise comparison of disease specific survivals of *PIK3CA's* regulatory SNP genotype TT, CT and CC groups in METABRIC set (p-values calculated by log-rank test if curves do not cross and by Two-Stage procedure if curves cross).

	ER positive	ER negative	PR positive	PR negative	HER2 positive	HER2 negative
TT : CT	0.689	0.077	0.516	0.745	—	0.869
TT : CC	0.852	0.0018	0.9306	2.57E-05	—	0.757
CT : CC	0.792	0.6	0.488	0.513	0.555	0.6271

Next, I looked for the possibility of the genotype influencing survival for specific characteristics of tumours. For that, I first divided the samples according to receptor statuses and then applied the survival analysis. I observed a significant difference in overall and disease specific survival for PR negative group (p-value = 0.03 and 0.0071, respectively), with the significant difference being between TT (median survival time 2.44 years for overall and disease specific survivals) and CC (median survival time 7.38 and for overall and disease specific survival, and not reaching median survival when assessing disease specific survival) genotypes group, with TT group having worse outcome (p-value = 0.00002). When assessing pairwise comparison between survival curves, I also observed a significant difference between TT (median survival time 2.44 years for overall and disease specific survivals) and CC (median survival time 3.76 and 4.83 years for overall and disease specific survivals, respectively) genotypes groups for ER negative tumours for both overall and disease specific survival (p = 0.002 and 0.0018, respectively; Figure 4.15, Table 4.7 and Annex L). In order to study the influence of both genotype and MADE categorization, I divided the tumours into MADE and mut=wt groups and then into the three genotype groups and performed the survival analysis. I did not observe any significant difference (Annex M).

Since the allele T was predicted to be associated with *PIK3CA's* higher expression and we hypothesise that this is associated with worse outcome, I put together the samples with T allele (TT and CT) and performed the survival analysis again comparing TT/CT genotype group with CC genotype group. The only significant difference observed was for disease specific survival for PR negative group, with TT/CT group surviving less than CC group TT/CT group with 4.2 years of median survival and the CC group having more than 50% of people alive at the end of observation period (p-value = 0.033; Figure 4.16 and Annex L).

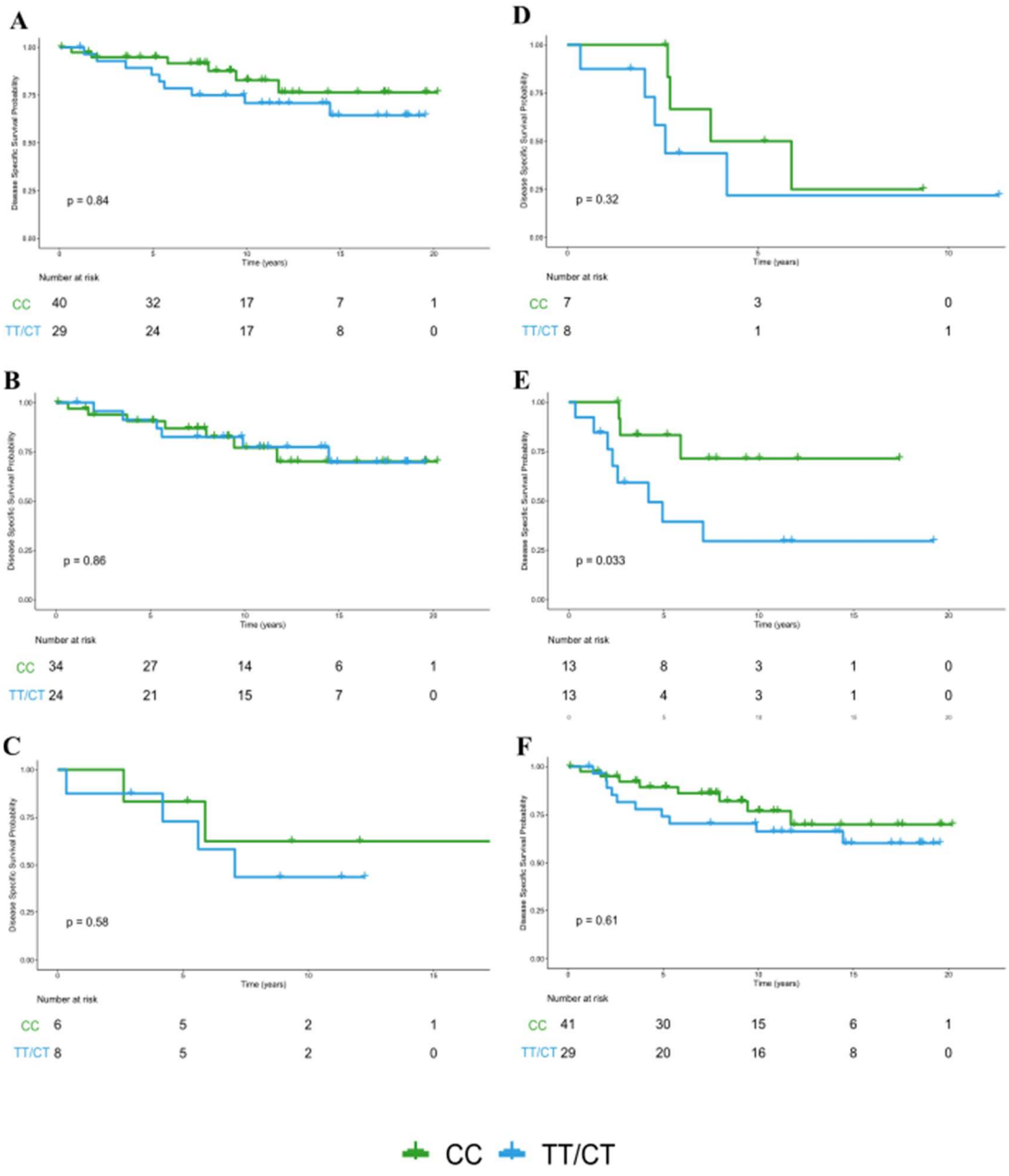


Figure 4.16: Kaplan-Meier analysis of disease specific survival of *PIK3CA's* regulatory SNP genotype TT/CT and CC groups in METABRIC set according to ER, PR and HER2 statuses. Disease specific survival in (A) ER positive group, (B) PR positive group, (C) HER2 positive group, (D) ER negative group, (E) PR negative group and (F) HER2 negative group

4.3 Mutant allele differential expression analysis of *TP53* mutations in breast cancer

4.3.1 Protein consequence of mutations

In view of my discoveries for cis-regulation of *PIK3CA* mutations in breast cancer, I set out to carry a similar analysis for *TP53*, the second most mutated gene in both METABRIC and TCGA datasets. First, I analysed the consequence of the *TP53's* somatic mutations present in both METABRIC and TCGA sets at the protein level. For that, I manually checked the OncoKB (Chakravarty *et al.*, 2017) for classification of the amino acid alteration and then analysed if there was a difference in the distribution of the type of mutations' consequence between different MADE categories. I found that the majority of mutations were likely oncogenic and there was a significant difference in the distribution of type of impact between MADE categories (p-value = 0.012 and 1.8e-06, for METABRIC and TCGA sets, respectively, Table Annex N). Nevertheless, there was no significant difference between the distribution of the "likely oncogenic" and the "oncogenic" subgroups. However, a significant difference was detected between these two groups and the "unknown" subgroup (Annex N). When looking at the distribution of the mutations on the p53 protein, I observed that some mutations occurred in the known hotspots, and others scattered throughout the gene (Figure 4.17), as expected due to the dual role of tumour-suppressor gene and oncogene.

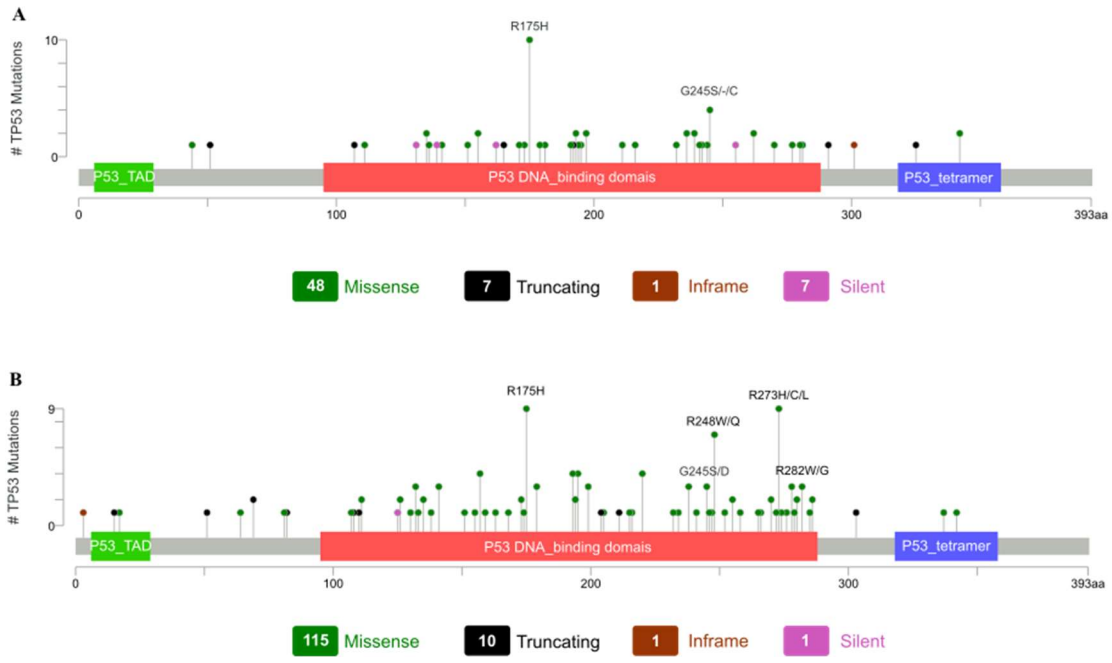


Figure 4.17: Amino acid alterations in p53 protein. The distribution of mutations in p53 for METABRIC set (A) and in TCGA set (B). The height of the lollipop represents the total number of each mutation. Only the amino acid changes in the known hotspots are written.

4.3.2 Differential Allelic Expression of *TP53*'s somatic mutations in Breast Cancer

Next, I set out to assess whether *TP53* somatic mutations would also have their functional effects somewhat modified by imbalances in allelic expression. In order to perform this analysis, mutant vs wild-type allelic expression analysis was carried for breast tumour samples carrying somatic *TP53* mutations, on two independent sets of data, the METABRIC (METABRIC set, n=63) and the TCGA (TCGA set, n=127) projects. The summary description of the two sets can be found in Annex O.

When looking at the distribution of α, β and γ ratios of *TP53's* somatic mutations in breast cancer tumours in both datasets, I observed that the average γ ratio (1.6963 and 1.9234, for METABRIC and TCGA sets respectively) was significantly higher than the average of the β ratios (pairwise Mann-Whitney U test p-value = 1.10e-10 and < 2e-16, for METABRIC and TCGA sets, respectively), but it was not significantly different from the α ratios (pairwise chi-square test p-value = 0.41 and 0.29, for METABRIC and TCGA sets, respectively, Figure 4.18, and Annex P)

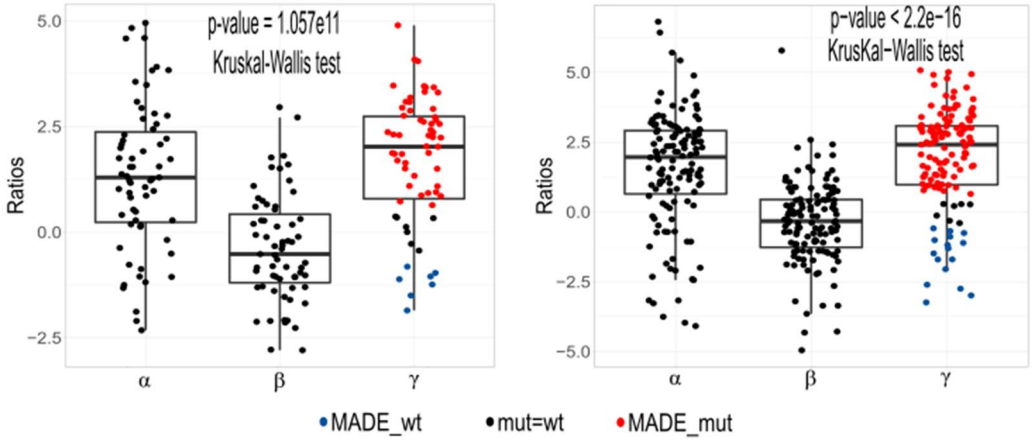


Figure 4.18: Distribution of *TP53*'s α, β and γ ratios in breast tumours in (A) METABRIC set and (B) TCGA set. Average γ was significantly higher than average β.

The majority of tumours, both in the METABRIC set (82.54%) and in the TCGA set (91.34%), displayed α ratios (total mutant vs wild-type allelic expression imbalances) greater than 1.5-fold. The same was true for γ ratios (the normalized expression allelic mutant/wild-type ratios) which showed a small increase in the percentages when compared to α ratios (88.89% and 94.49%, for METABRIC and TCGA sets, respectively). When looking at the

proportion of samples with allelic imbalance, I observed, in both datasets, that the majority expressed more of the mutant allele (Table 4.8).

When looking at expression imbalances values distribution in both sets, tumours with preferential expression of the mutant allele displayed higher fold changes differences than tumours with higher expression of the wild-type allele. This was observed for α ratios, but also for γ ratios (Table 4.8).

Table 4.8: Summary of allelic imbalance of *TP53's* somatic mutations in breast tumours

	METABRIC set (n=63)		TCGA set (n=127)	
	α	γ	α	γ
Mutation allelic imbalance				
absolute ratio ≥ 0.58 (%)	82.54	88.89	91.34	94.49
Mut Allele > WT Allele (%)	66.66	77.77	74.8	81.9
WT Allele > Mut Allele (%)	15.87	11.11	16.5	12.6
Maximum Fold Difference				
Mutant Allele	31	29.81	112.5	33.73
Wild-type Allele	5	3.61	17	9.46

4.3.3 Positive selection for cis-regulated preferentially expressed mutated alleles

In order to dissect the contribution of both cis-regulation and chromosomal copy number level to the differences in mutant allelic expression, I investigated whether there was positive, neutral or negative selection for the mutant allele imbalance, by comparing *TP53's* γ and β values for each tumour, in both sets. I found that, in both cohorts, the majority of samples (54.46% and 49.6% for the METABRIC and TCGA sets, respectively) had positive γ and negative β values, indicating positive selection for cis-regulation, and not for allele count (Figure 4.19). These samples showed preferential expression of the mutant allele, albeit the larger proportion of the wild-type allele at the DNA level.

The fraction of samples that displayed preferential allelic expression and higher allele copy number of the mutation was the second largest subgroup for both datasets (27.69% and 35.43% for the METABRIC and TCGA sets, respectively). To evaluate how much of these results might be confounded by different levels of cellularity or tumour purity, I introduced this information in the analysis for the METABRIC set and found no association between these parameters and the β or γ values (Figure 4.20). I could not perform the same analysis for the

TCGA set for lack of information on cellularity. Altogether, these results indicate that cis-regulation of expression of the mutant allele is positively selected for in tumours, and that it is predominant to allelic copy number variation in determining preferential expression of mutations in tumours.

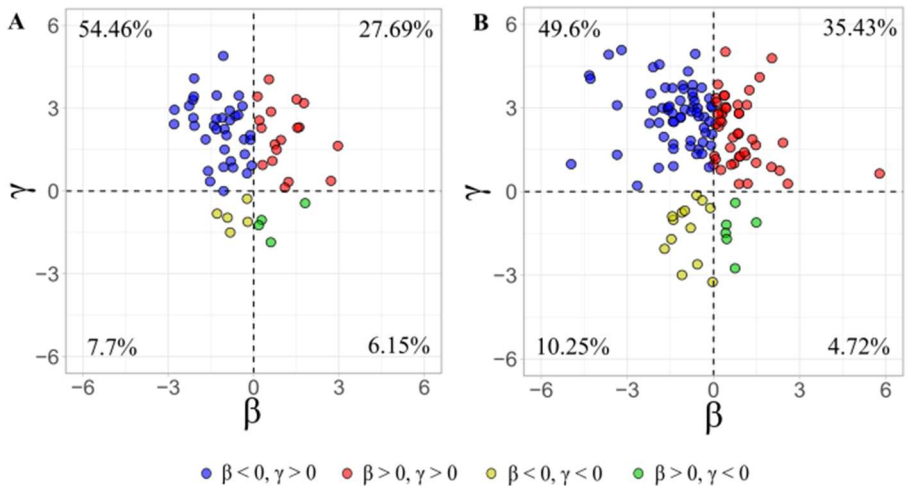


Figure 4.19: Comparison of matched *TP53's* β and γ values showing predominance of tumours showing preferential allelic expression of the mutated allele, albeit higher number of wild-type allele copies.

Since for the *TP53's* analysis I did not select one specific type of mutation, instead I analyzed the distribution of the mutations type when comparing matched β and γ values. I observed that the majority of missense mutations, when taking into account cis-regulation, are preferentially expressed and nonsense mutations are not, for both datasets (Figure 4.21 and Figure 4.22).

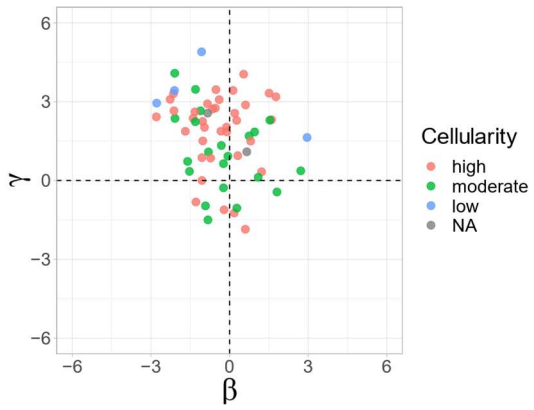


Figure 4.20: Comparison of matched *TP53's* β and γ values for METABRIC set, with dots coloured by cellularity levels showing that there is no pattern of association between ratios and cellularity

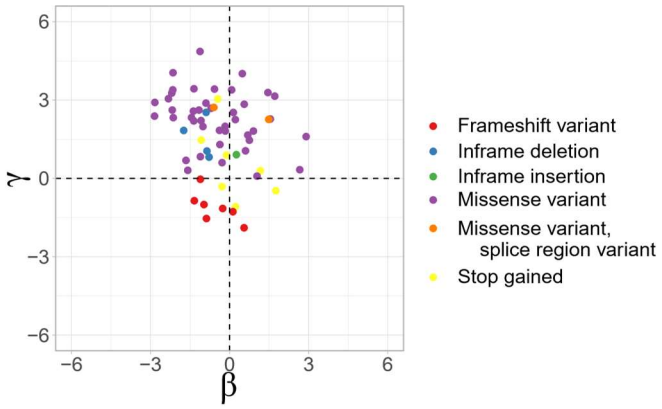


Figure 4.21: Comparison of matched *TP53's* β and γ values for METABRIC set, with dots coloured by type of mutation showing that the majority of missense mutations preferentially express the mutated allele.

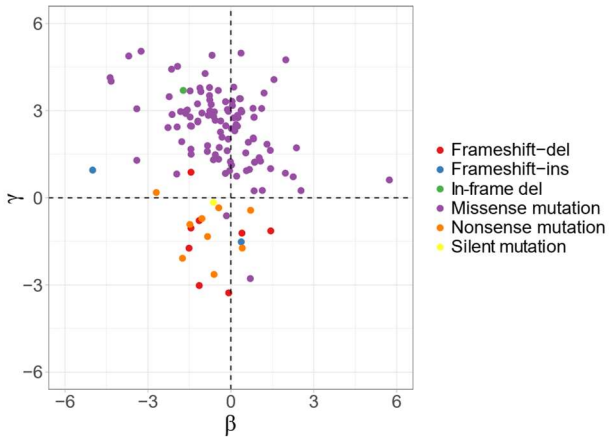


Figure 4.22: Comparison of matched *TP53's* β and γ values for TCGA set, with dots coloured by type of mutation showing that the majority of missense mutations preferentially express the mutated allele.

4.3.4 Clinical correlation of mutant *TP53's* expression ratios

Next, the clinical impact of mutant *TP53*'s allelic imbalance on tumour biology was explored by correlating the mutant/wild-type expression ratios with breast cancer clinicopathological characteristics, such as the hormonal receptor and HER2 statuses, PAM50 and Integrative Clusters classifications, tumour size, stage, histological grade, lymph node status and patients age, for the METABRIC and TCGA sets. To accomplish this, first, the allelic expression ratios were used as a continuous variable and applied a bivariate analysis, whose main results are shown in Figure 4.23, Figure 4.24 and Annex Q.

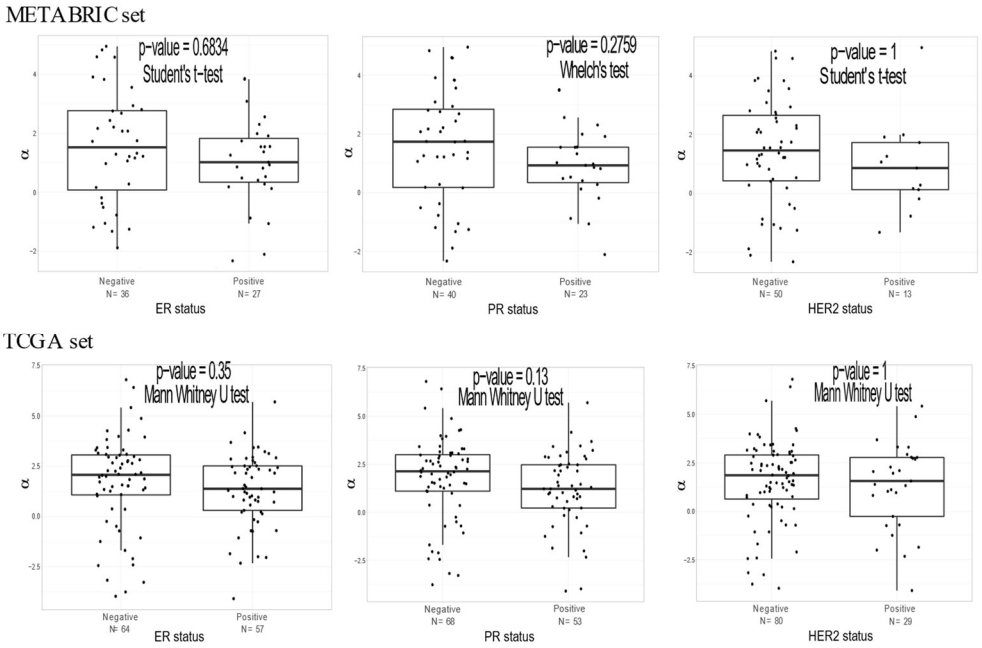


Figure 4.23: Association of *TP53's* α ratios and receptors statuses. Upper Line: Boxplots showing association between α ratios and ER, PR and HER2 statuses, respectively for the METABRIC set. Bottom line: Boxplots showing association between α ratios and ER, PR and HER2 statuses, respectively for the TCGA set. Each dot represents a sample.

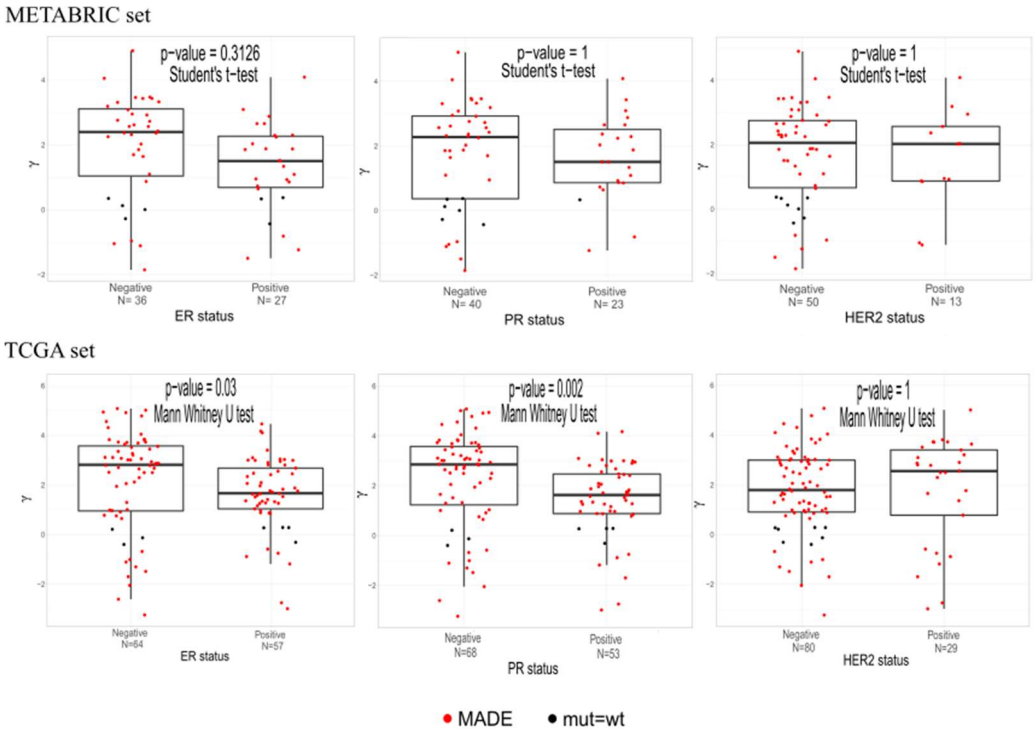


Figure 4.24: Association of *TP53's* γ ratios and receptors statuses. Upper line: Boxplots showing association between γ ratios and ER, PR and HER2 statuses for the METABRIC set. Bottom line: Boxplots showing association between γ ratios and ER, PR and HER2 statuses for the TCGA set. Each dot is a sample. Red dots correspond to tumours with MADE, while black dots represent samples with no allelic imbalance.

There was no association between α and γ ratios and the clinicopathological characteristics for the METABRIC set. The analysis of the TCGA set showed association between γ ratio and ER status (p-value = 0.03) and PR status (p-value = 0.002), but not with other characteristics or with α ratios.

To study if the distribution of patient age and tumours clinicopathological characteristics were different between MADE groups, chi-square tests were performed for the METABRIC and for the TCGA sets. There was no significant difference in distribution of grade, size, stage, integrative clusters, PAM50 classification, number of positive lymph nodes, age ER status or HER2 status between different MADE groups for neither datasets (Annex R).

4.3.5 Mutant *TP53* expression ratios and clinical outcome

Then, I set out to investigate if there was an impact of differential cis-regulation of *TP53*'s mutations on clinical outcome. In order to achieve this, I performed univariate survival analysis comparing the *TP53's* MADE categorization with respect to overall survival and to disease specific survival. I found that there was no association between MADE and overall or disease specific survival for METABRIC set (Figure 4.25, Table 4.9 and Annex S).

For TCGA set, there was difference between MADE and mut=wt groups (p-value = 0.0265) for the overall survival, with median survival of 6 years for MADE group and 4.26 years for mut=wt group (Figure 4.26, Table 4.10 and Annex U). When further stratifying MADE group and analysing overall survival, I found that MADE_mut group had a significant poorer survival when compared to mut=wt group (p-value = 0.0264), but there was no difference between the others. I found no significant difference when assessing disease specific survival in TCGA set.

Kaplan-Meier analysis was also performed for α ratios in both datasets. There was no statistically significant difference between α_DAE and α_noDAE groups for the METABRIC and TCGA sets when considering overall survival (p-value = 0.99 and 0.63, respectively), nor when taking into account disease specific survival (p-value = 0.64 and 0.35, respectively, Annexes T and V). I also found no significant difference when I further stratified α_DAE group and analysed overall and disease specific survivals for both datasets.

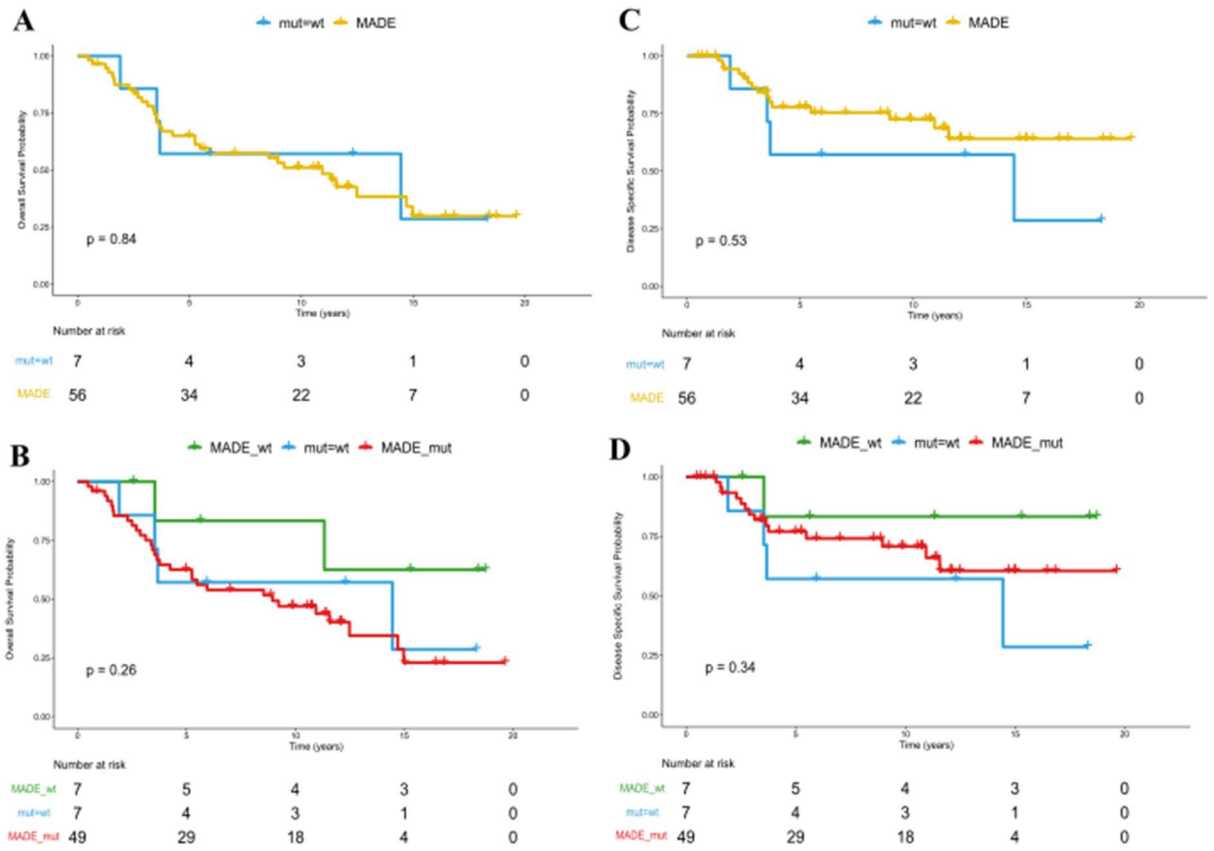


Figure 4.25: Kaplan-Meier analysis of overall survival and disease specific survival of the *TP53's* in the METABRIC data set. (A) Overall Survival of METABRIC set divided in MADE and mut=wt groups. (B) Overall Survival with MADE group further divided in tumours that express more wild-type allele (MADE_wt) and tumours that express more mutated allele (MADE_mut). (C) Disease Specific Survival of METABRIC set divided in MADE and mut=wt groups. (D) Disease Specific Survival with MADE group further divided in tumours that express more wild-type allele (MADE_wt) and tumours that express more mutated allele (MADE_mut).

Table 4.9: P-values of pairwise comparison of overall and disease specific survivals of *TP53's* MADE groups in METABRIC set (p-values calculated by log-rank test if curves do not cross and by Two-Stage procedure if curves cross)

	OS	DSS
MADE_wt : mut=wt	0.3014	0.1391
MADE_wt : MADE_mut	0.8509	0.7565
mut=wt : MADE_mut	0.8746	0.3555

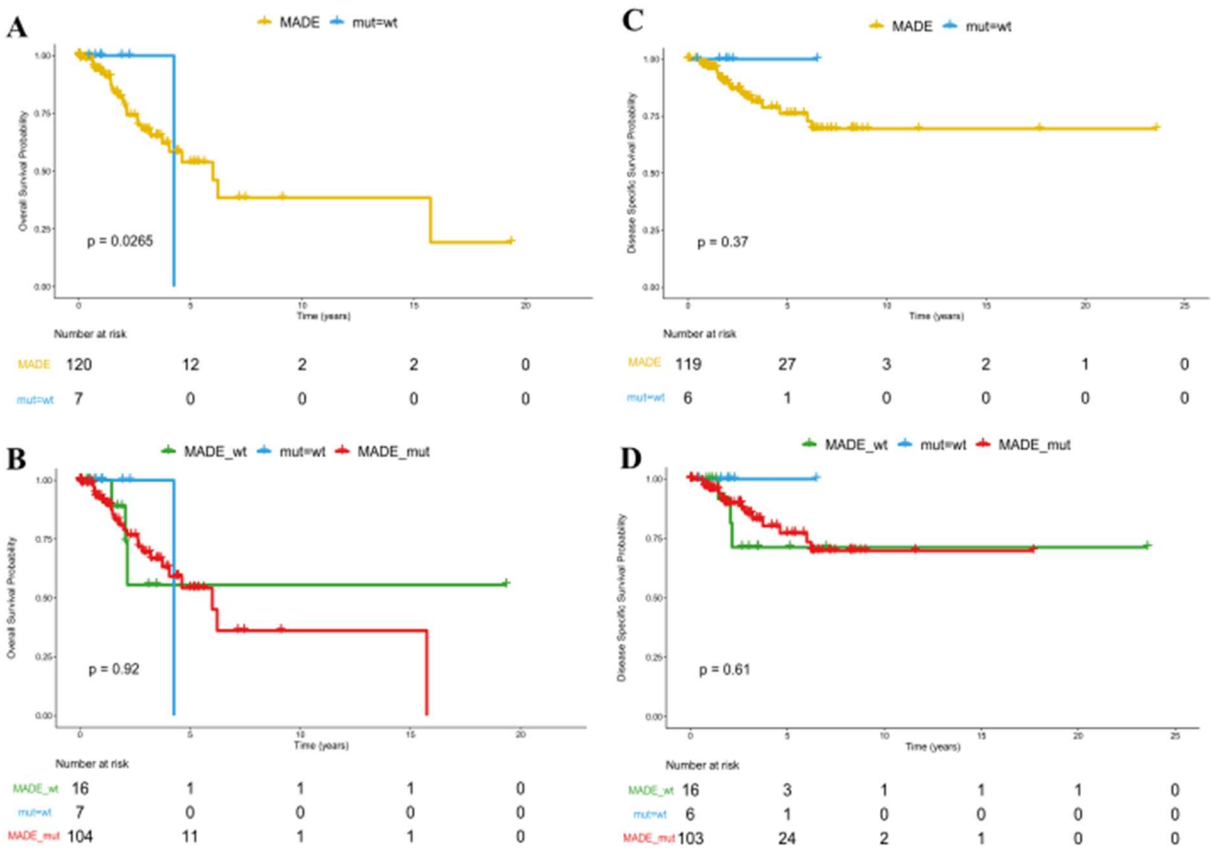


Figure 4.26: Kaplan-Meier analysis of overall survival and disease specific survival of *TP53's* MADE in the TCGA data set. (A) Overall Survival of TCGA set divided in MADE and mut=wt groups. (B) Overall Survival with MADE group further divided in tumours that express more wild-type allele (MADE_wt) and tumours that express more mutated allele (MADE_mut). (C) Disease Specific Survival of TCGA set divided in MADE and mut=wt groups. (D) Disease Specific Survival with MADE group further divided in tumours that express more wild-type allele (MADE_wt) and tumours that express more mutated allele (MADE_mut).

Table 4.10: P-values of pairwise comparison of overall and disease specific survivals of *TP53's* MADE groups in TCGA set (p-values calculated by log-rank test if curves do not cross and by Two-Stage procedure if curves cross)

	OS	DSS
MADE_wt : mut=wt	0.3933	0.3466
MADE_wt : MADE_mut	0.9055	0.9485
mut=wt : MADE_mut	0.0264	0.3823

Next, I explored the effect of γ ratios in survival according to ER, PR and HER2 statuses in order to study a possible different impact of *TP53's* allelic imbalance in the different subgroups of tumours. First, the datasets were divided into receptor positive and receptor negative groups, then Kaplan-Meier overall and disease specific survival analyses were applied. I found no difference in METABRIC set overall or disease specific survivals (Annex T). For

the TCGA set, I found that there was a significant overall survival difference between MADE and mut=wt groups, with MADE having worse survival, for both ER positive and PR negative groups (p-value = 0.043 and 0.042, respectively, Figure 4.27). When further stratifying MADE group in TCGA set, there was a significant difference between mut=wt and MADE_mut groups' overall survival for ER positive tumours (p-value = 0.042), with MADE_mut with poorer outcome, and between MADE_wt and mut=wt groups' overall survival for HER2 negative tumours (p-value = 0.025), with MADE_wt presenting poorer outcome. There was no significant difference for TCGA set when assessing disease specific survival (Annex V).

When assessing survival for the subgroups divided by receptors status and considering α ratios, there was no statistically significant difference in overall, nor disease specific survival in TCGA set (Annex V). For METABRIC set, I found significant overall survival difference between α_DAEwt and α_noDAE groups in HER2 positive tumours (p-value = 0.04), with α_DAEwt group presenting better outcome, and between α_DAEwt and α_DAEmut groups in HER2 negative tumours (p-value = 0.045), with α_DAEmut presenting poorer outcome (Figure 4.28 and Table 4.11), and found no significant difference when assessing disease specific survival.

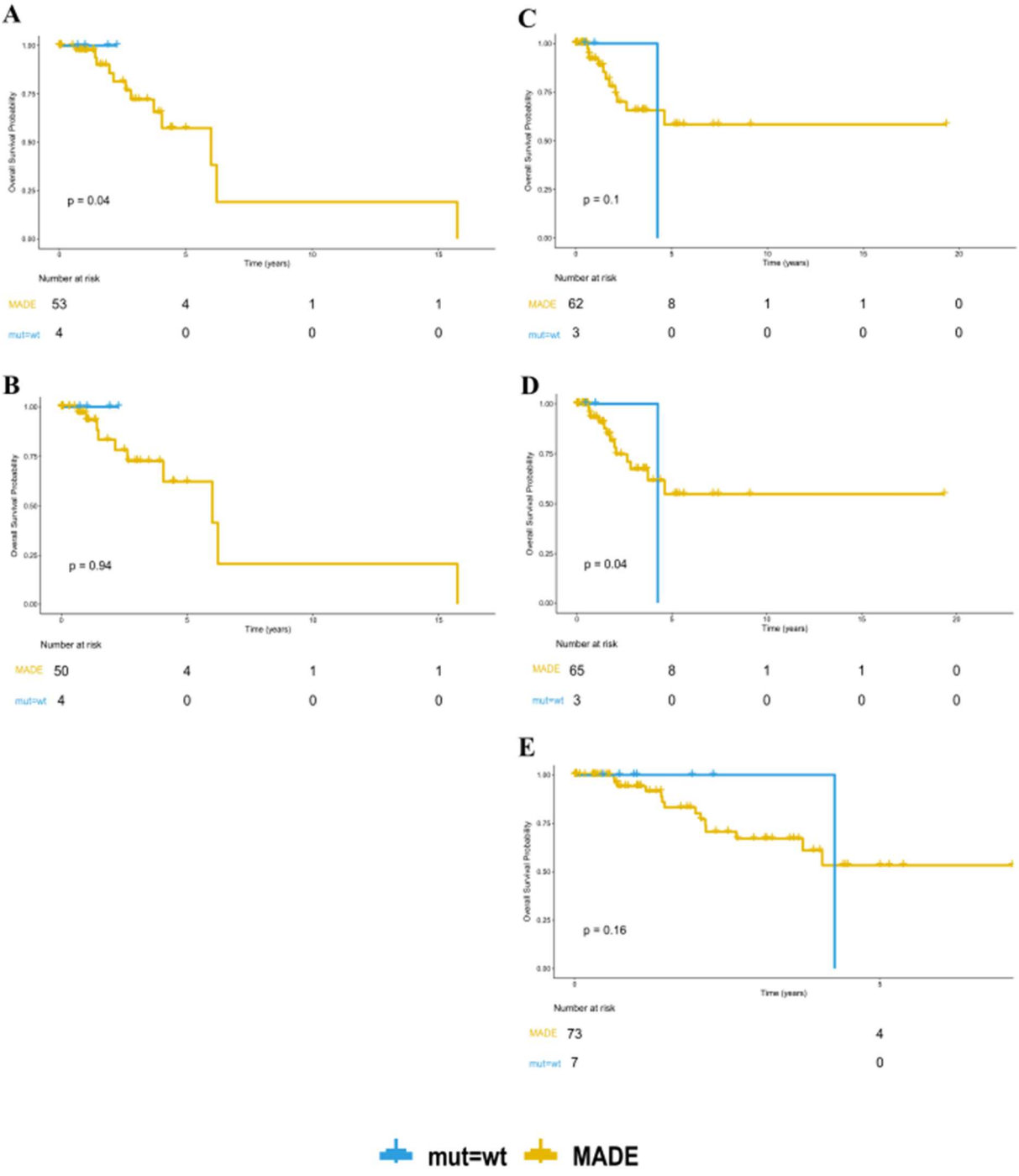


Figure 4.27: Kaplan-Meier analysis of overall survival according to ER, PR and HER2 statuses of *TP53's* MADE in the TCGA data set. Overall in (A) ER positive group, (B) PR positive group, (C) HER2 positive group, (D) ER negative group, (E) PR negative group and (F) HER2 negative group.

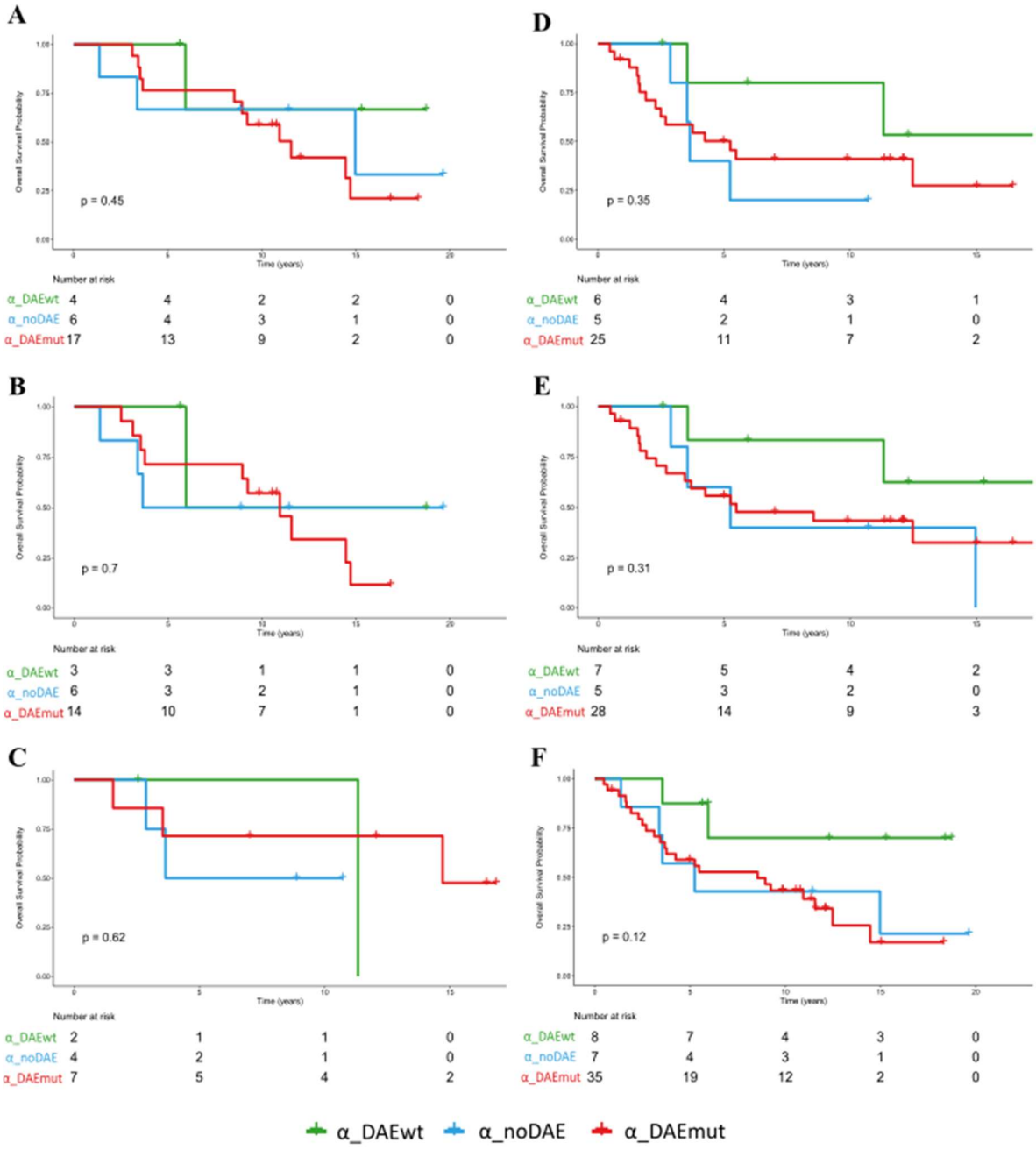


Figure 4.28: Overall survival of *TP53's* α_DAE groups in METABRIC set according to ER, PR and HER2 statuses. Overall survival in (A) ER positive group, (B) PR positive group, (C) HER2 positive group, (D) ER negative group, (E) PR negative group and (F) HER2 negative group.

Table 4.11: P-values of pairwise comparison of overall survival of *TP53's* α_DAE groups in METABRIC set according to ER, PR and HER2 statuses (p-values calculated by log-rank test if curves do not cross and by Two-Stage procedure if curves cross)

	ER positive	ER negative	PR positive	PR negative	HER2 positive	HER2 negative
α_DAEwt : α_noDAE	0.4978	0.0796	0.3397	0.0983	0.0402	0.0808
α_DAEwt : α_DAEmut	0.572	0.2224	0.9456	0.1724	0.7442	0.0447
α_noDAE : α_DAEmut	0.1109	0.0611	0.21132	0.1011	0.5062	0.6864

CHAPTER 5
DISCUSSION

5 Discussion

In the last decade, many efforts have been concentrated to supplement the morphological classification of breast carcinoma with molecular parameters that can provide a clearer appreciation for the heterogeneity of breast cancer and for better prediction of tumour behaviour, to improve therapeutic strategies (Makki, 2015).

Changes in copy number have been extensively exploited to the level of helping to stratify the patient population, which in the METABRIC project led to the definition of ten subtypes based on the integration of genomic and transcriptomic profiles of breast tumours (Curtis *et al.*, 2012). Other studies have looked specifically at allelic bias in the copy number balance between mutant and wild-type alleles, ranging from small to large imbalances, and found that it has prognostic value and also predicts response to treatment (Mitsudomi *et al.*, 1991; Soh *et al.*, 2009; Oakley and Chiosea, 2011; Bielski *et al.*, 2018; Stagni *et al.*, 2018). Bielski *et al.* also demonstrated that although allelic imbalance was not driven by the presence of a mutant allele in these cancers, it subsequently occurred to produce a mutant allele dosage increase, likely providing a fitness advantage to the evolving malignant clone (Bielski *et al.*, 2018).

Nevertheless, all studies on allelic imbalance have been done at DNA level. But imbalances can also be generated by control of gene expression, more specifically by cis-regulatory variants, which control gene expression at an allele-specific level. These variants have been shown to play a role in mutation penetrance (Castel *et al.*, 2018), particularly in the context of germline mutation carriers (Cox *et al.*, 2011; Maia *et al.*, 2012). However, how they impact on tumour biology and clinical outcome is still under-explored, and is the central aim of this thesis.

I chose to study the allelic imbalance of somatic mutations through the analysis of the differential allelic expression in tumours with heterozygous mutations. Differential allelic expression is the most robust approach to detect the effect of and map cis-regulatory variation. It focuses on the cis-effect alone, because it allows for the expression of one allele to serve as the internal control for the other, thus eliminating the effect of trans-factors, as both alleles are under the influence of the same trans-regulatory elements (Pastinen and Hudson, 2004; Xiao and Scott, 2011).

It is known that the proportion of transcribed variants tend to vary across different RNA read depths, with low number read depth having a lower power to detect rare variants, making it necessary an appropriate minimum read depth cut-off in RNAseq for differential allelic expression analysis. There is evidence that a read depth of 10 is a reliable cut-off to detect variants in RNAseq (The Cancer Genome Atas Research, 2013; Rhee *et al.*, 2017; Batcha *et al.*, 2019). Nevertheless, as the objective was a quantitative one, rather than qualitative (e.g. calling genotypes), I chose to use the cut-off for RNAseq and DNAseq depth of 30 reads, which has been used in our group in other studies, and has shown that albeit conservative, has been a reliable one.

When assessing the allelic imbalance of mutated genes in two large breast cancer cohorts, METABRIC and TCGA projects, I observed that almost all somatically mutated genes had differential allelic expression. The only gene with no allelic imbalance observed in the METABRIC set was *ERBB3*, and this probably was due to *ERBB3* mutation being present in only two samples, since in the TCGA set, this gene presented allelic imbalance. The two genes mutated in more samples, in both sets, were *PIK3CA* and *TP53*, which are genes previously implicated in breast cancer development (Stephens *et al.*, 2012). Thus, I chose to study *PIK3CA* and *TP53's* somatic mutations differential allelic expression.

5.1 Mutant allele differential expression analysis of *PIK3CA* mutations in breast cancer

First, I observed that, for both METABRIC and TCGA sets, the majority of *PIK3CA's* mutations occurred on the known hotspots E542K, E545K and H1047R (Bachelot *et al.*, 2011), which means that the datasets used for the analysis were not biased and reflected the usual mutations distribution that occurs in the patient population.

Next, I showed that *PIK3CA's* mutations displayed differential allelic expression, in both datasets, with approximately half of the tumours with *PIK3CA's* mutation displaying allelic imbalance due to cis-regulation. When comparing overall expression with cis-regulation effect alone (i.e. comparing differential allelic expression at the α ratios level with differential allelic expression at the γ ratios level), I observed that most of the allelic imbalance observed at RNA level is due to cis-regulation. When looking at the overall expression (α ratios) the majority of the tumours with differential allelic expression expressed more of the wild-type allele, but when adjusting for copy number alterations and looking for cis-regulation effect (γ

ratios) I observed that most tumours with differential allelic expression expressed more of the mutated allele.

I also found that there is a positive selection for preferential expression of the somatic mutant allele in the analysed tumours. This was suggested by the comparison of the matched β and γ ratios, which showed that although the majority of tumours displayed more copies of wild-type allele at the DNA level ($\beta < 0$), they expressed more of the mutated allele ($\gamma > 0$) due to cis-regulation. This result places cis-regulation at the forefront of tumour evolution for the first time.

Since I observed higher copies of wild-type allele at the DNA level, and this could be due to higher proportion of normal cells in the sample (the RNAseq data was collected from bulk tumour), I analysed the cellularity of the tumours (higher cellularity meaning higher proportion of tumour cells). I did not observe any typical pattern of distribution of cellularity, and therefore concluded that it was not influencing the observed result.

Furthermore, I show that the preferential expression due to cis-regulation of *PIK3CA's* mutated allele is associated with PR negative tumours in both METABRIC and TCGA sets, and with HER2 positive in the METABRIC set and with ER negative in the TCGA set. It is known that ER and PR positive breast cancers have better prognosis than ER and PR negative ones (Dunnwald, Rossing and Li, 2007), and that amplification and/or overexpression of the HER2 gene in breast cancer is associated with adverse prognosis (Chia *et al.*, 2008). Thus, the preferential expression of *PIK3CA's* mutated allele is associated with tumours with worse prognosis. Nevertheless, the possibility of causality was not assessed by this study.

Clinical therapy-selection biomarkers often assay mutations using DNA as an analyte, such as KRAS assays designed to identify responders to anti-EGFR monoclonal antibody therapy. However, if the wild-type allele is selectively transcribed, the mutation may not have therapeutic impact and the merit of using a DNA-based assay for clinical decision-making may be problematic (Castle *et al.*, 2014). A recent study revealed that even modest dosage changes, resulting in allelic imbalance due to copy number alterations, can drive clonal outgrowth and modulate drug responses (Bielski *et al.*, 2018). The differences in the expression of the mutant allele observed in this study could possibly explain why, even though the frequency of the *PIK3CA* mutation is high in breast cancers, the response to PI3K inhibitor therapy has not been

borne out in clinical studies so far, and the prognostic significance of detecting somatic *PIK3CA* mutation in a breast tumour is uncertain (Keegan *et al.*, 2018).

The univariate survival analysis of the METABRIC set has shown that patients with γ ratios allelic imbalance (MADE group) had worse overall and disease specific survivals than those without the imbalance (mut=wt group). Interestingly, when further dividing MADE tumours that express more of the wild-type allele (MADE_wt) and those that express more of the mutated allele (MADE_mut), the group with the poorer survival was the MADE_wt group. One possible explanation for the group that express less mutation to have a worse outcome is that these tumours could have a combination of other mutations that would contribute for the worse prognosis and even though there is a mutation in *PIK3CA*, it is not a driver for those tumours. Nevertheless, this possibility was not assessed here. One limitation of this analysis is that although I observed that MADE_wt group had poorer outcome, the group is comprised of only six samples and, since I cannot increase sample number, I cannot discard the fact that the results could be due to chance.

The multivariate analysis of the METABRIC set did not show statistically significant association between γ ratios and overall or disease specific survivals, but since it also does not show association with PR and HER2 statuses, which have known impact on survival (Dunnwald, Rossing and Li, 2007; Chia *et al.*, 2008), I cannot reject the hypothesis that the study may not have the power necessary to detect the effect of the γ ratios on survival. To corroborate this, the power of a survival analysis was related to the number of events rather than the number of participants and a simulation work has suggested that at least 10 events need to be observed for each covariate considered in the multivariate analysis, and anything less would lead to problems, for example, the regression coefficients would become biased (Peduzzi *et al.*, 1995). In my regression model I included patients age, tumour's histological grade and stage, ER, PR and HER2 statuses, which are known independent prognostic factors (Sun *et al.*, 2016), and I had approximately six events per covariate, which could be a low number of events impairing the results.

The survival analysis of the MADE compared to mut=wt groups for TCGA set was not statistically significant and this could be due to no association between γ ratios and overall and disease specific survival or because the sample was mainly composed of Luminal A breast tumours (~68% of samples) and follow up time was short (median 2.09 years).

There are five primary intrinsic subtypes of breast cancers, including the most aggressive basal-like subtype, which commonly recurs within a few years, and the least aggressive Luminal A subtype, which is a highly heterogeneous group, in which the prognosis

of each patient is extremely variable. More than half of all disease recurrences in Luminal subtypes of breast cancer occur six years or more after diagnosis, particularly following five years of adjuvant anti-oestrogen therapy (Lim, Filho and Winer, 2012; Sun *et al.*, 2016; Liu *et al.*, 2018).

Time to event studies, such as survival analysis, must have sufficient follow-up to capture enough events and thereby ensure there is sufficient power to perform appropriate statistical tests. The proposed length of follow-up for a prospective study will be based primarily on the severity of the disease or prognosis of the participants. For example, for a lung cancer trial a 5-year follow-up would be more than adequate, but this follow-up duration will only give a short- to medium-term indication of survivorship among breast cancer patients, specially hormone receptor positive tumours (Luminal A tumours are mainly hormone receptor positive). An indicator of length of follow-up is the median follow-up time (Clark *et al.*, 2003), which for the TCGA set is around two years, as said above.

Corroborating this observation, an integrated analysis of the TCGA clinical data, showed that for less aggressive cancer types like breast cancer, appropriate use of overall survival depends on its subtype (Liu *et al.*, 2018). Hence, given a relatively short follow-up time of the TCGA data set and the fact that tumours are mainly subtype Luminal A, which has a latter recurrence, the survival analysis of the validation set could lack statistical power.

Interestingly, when analysing α ratios disease specific survival for the TCGA set, I observed that α_DAE group had better survival than α_noDAE group and that this was mainly due to better survival of α_DAEwt group (which preferentially express wild-type allele) when compared with α_noDAE group. This was according to the expectation that the preferential expression of the wild-type allele would be beneficial to the patient, and the result was probably observed due to the larger number of patients in the α_DAEwt group.

I also observed associations between ER positive, PR positive and HER2 positive groups and MADE, when MADE was further stratified into MADE_mut and MADE_wt, for both overall and disease specific survivals for the METABRIC set. I observed, again, that MADE_wt group was the one with poorer survival, but this conclusion needs to be taken with caution due to small number of subjects in this group. When applying pairwise comparison for overall survival, PR negative MADE_wt and MADE_mut groups were also significantly different, which corroborates the apparently significance of *PIK3CA's* somatic mutations allelic imbalance especially for PR status. However, MADE_wt group in PR negative group was

comprised with only one sample, so the result should be taken even more carefully. The survival analysis of MADE groups for ER negative and HER2 positive were not significant, and one possible explanation is the small number of subjects that have tumours ER negative and HER2 positive (16 and 15, respectively).

As our group had data suggesting that rs2699887 is a cis-regulatory SNP of *PIK3CA* in normal breast tissue, with the T allele (minor allele) associated with higher expression of *PIK3CA*, I also analysed its implications in γ ratios in tumours. I analysed rs2699887 genotype in the METABRIC set of breast cancer samples and observed no difference of *PIK3CA's* somatic mutation allelic imbalance occurrence between the 3 possible genotypes (TT, CT, CC). It was expected that more allelic imbalance would be detected in the CT genotype than in the others, if rs2699887 were the main cis-regulatory variant. Next, I assessed the impact of rs2699887 genotype on patients' survival and the only significant results observed were in PR negative and ER negative subgroups with TT genotype group presenting poorer survival.

Support for the clinical significance of rs2699887 comes from a recent association with worse overall survival in melanoma (Morgese *et al.*, 2017) and with higher risk of brain metastasis in non-small cell lung cancer, likely because of an increase in PI3K signalling (Li *et al.*, 2013). However, one other study on endometrial carcinoma found the heterozygous genotype was associated with better survival (Wang *et al.*, 2012), and another on lung cancer found the alternative allele associated with higher risk of toxicity (Pu *et al.*, 2011). So, its clinical role is still uncertain.

My results suggest that albeit in normal tissue rs2699887 might be the causal variant of differential allelic expression, in tumours other somatic events might be contributing towards the allelic imbalances in expression that I detected. It has been shown that non-coding regulatory somatic mutations occur as frequently as coding somatic mutations in breast cancer, influencing the expression of cancer driver genes (Rheinbay *et al.*, 2017). There is also evidence that somatic mutations in the regulatory elements of *ESR1* (gene that code oestrogen receptor α, a driver in some breast cancers) were identified. Also, pan-cancer analyses have reported positive selection in mutated regulatory elements proximal to known cancer related genes (Melton *et al.*, 2015). Alterations in epigenetic marks can also alter gene expression and have allele-specific effects in tumour development (Gonzalez-Zulueta *et al.*, 1995). In my study I studied the effect of allelic imbalance in somatic coding mutations, and not the mechanism behind it, which warrants an investigation of its own. Therefore, I cannot rule out the possibility

that the differential allelic expression that I observed in *PIK3CA's* somatic mutated breast cancers are due to other mutations and alterations in cis-regulatory elements. This could be one explanation why I did not find a strong association with normal cis-regulatory variant in rs2699887.

5.2 Mutant allele differential expression analysis of *TP53* mutations in breast cancer

In view of the clinical significance that I identified in the study of the preferential expression of *PIK3CA's* mutated allele, and the previous data from our group supporting the existence of cis-regulatory variants for *TP53* (Maia *et al.*, 2009), I set out to also analyse somatic mutation allelic imbalance in this gene. I did not observe differences between the distribution of likely oncogenic and oncogenic mutations and MADE group, and also the pattern of mutations on p53 protein was as expected for p53's known dual role, with some mutations occurring in the known hotpots (R175, G245, R248, R273 and R282), which is a characteristic of oncogenes, and other truncating and nonsense mutations scattered throughout the protein which is frequent in tumour suppressor genes (Liu, Zhang and Feng, 2013; Soussi and Wiman, 2015). Combined, these observations show that the datasets used for this analysis were, also, not biased and reflected the usual mutations distribution that occurs in the patient population.

Then, I observed preferential expression of *TP53's* mutated allele when looking into cis-regulation specifically (γ ratios), but differently from *PIK3CA's* analysis, I found that the allelic ratios from cis-regulation (γ ratios) were significantly higher than those from copy-number (β ratios) but not from the overall allelic ratios (α ratios) This means that there is general preferential expression of the mutant allele, and that this is mostly powered by cis-regulation. Another observation is that the proportion of tumours preferentially expressing the mutated allele, compared to that of tumours preferentially expressing the wild-type allele, is higher for both the overall and cis-regulation-specific ratios (α and γ ratios, respectively), at around 75% percent. This means that the mean difference observed between these ratios is due to many tumours presenting this characteristic, rather than being the effect of a few outliers with extreme ratios. However, when comparing copy-number and cis-regulation influence (β and γ ratios, respectively), of the samples preferentially expressing the mutated allele, one third also had more copies of the mutated allele, whilst for *PIK3CA* this proportion was around 6%. These results support that cis-regulation is one of the most important components influencing

mutation expression, but differently from *PIK3CA*, *TP53's* copy number alteration is also an important factor for the preferential expression of mutated allele.

Another interesting observation was that the majority of missense mutations preferentially expressed the mutated allele, while nonsense mutations did not. This could be due to the fact that the RNAseq does not quantify all nonsense mutations, because these mutations are mostly not expressed. Corroborating this, Rhee *et al.* demonstrated that tumours with nonsense and frameshift indels mutations preferentially express the wild-type allele, which they claim is consistent with a known biological phenomenon of nonsense-mediated decay where the transcripts containing those truncating mutations are relatively deficit compared to those with wild-type alleles. They have also shown across five tumour types, that 11.1–20.4% and 6.4–11.4% of missense mutations preferentially expressed mutated allele and wild-type-allele, respectively, suggesting that a substantial number of missense mutations are associated with allelic imbalance. Also, *TP53* mutant allele was found to be preferentially expressed in most tumour types, including breast cancer (Rhee *et al.*, 2017).

As for the *PIK3CA's* analysis, there was not a significant pattern of distribution of cellularity levels that could explain the higher copies of wild-type allele at the DNA level, for samples that expressed more *TP53's* mutated allele at γ ratios level. Thus, corroborating to the importance of cis-regulation.

When analysing the clinical effect of *TP53's* mutations, I observed a significant preferential expression of the mutated allele in ER negative and PR negative groups for TCGA set, when inspecting γ ratios. In the METABRIC set analysis, mean γ ratios were identical between ER, PR and HER2 groups, which could be due to lack of statistical power in our set, as there were only 63 tumours with *TP53's* somatic mutations.

When assessing survival of patients with *TP53's* mutations and its association with differential allelic expression, I observed that both METABRIC and TCGA sets had low number of tumours with no allelic imbalance (mut=wt groups), with both sets having only seven samples each in this group, which again limits the statistical power of the analysis. So, for the METABRIC set, I cannot discard that the non-significant results are due to low total sample number (63 samples), together with low number of samples in mut=wt group.

For the TCGA set, which is a larger set, I observed poorer survival of MADE_mut group when compared to mut=wt group, which suggests that preferential expression of *TP53's* mutated allele could be associated to worse outcome. Even though the mean follow-up was as

low as for *PIK3CA's* TCGA set, the fact that the majority of *TP53's* mutated tumours in this set were more aggressive basal-like and HER2-overexpressed subtypes (Sorlie *et al.*, 2001; Sun *et al.*, 2016; Liu *et al.*, 2018), this leads to more deaths in a shorter period. Therefore, there was higher power for this survival analysis, despite the short median follow-up time. The already mentioned lack of samples with no allelic imbalance (mut=wt group), was again a weakness of this analysis.

In spite of the association of *TP53's* MADE with poorer survival, I did not find the same for disease specific survival analysis. This could mean that other causes of death are more significant, but we cannot rule out that there were not enough disease specific deaths in the short period of observation. For example, the mut=wt group had seven subjects and only one death by cancer. Therefore, a validation set with longer follow up should be used to reanalyse this.

When dividing datasets according to receptor status, MADE group presented poorer survival in ER positive and PR negative groups. This, also, needs to be taken carefully due to even smaller number of mut=wt group and the fact that disease specific survival analysis was, again not significant.

When assessing overall mutation allelic ratios (α ratios), the α_noDAE group showed poorer overall survival when compared with α_DAEwt group, in HER2 positive tumours, but the sample set was so limited (n= 4 vs 2, respectively), that the biological impact of this is weak.. When looking at HER2 negative tumours and the same ratios, the α_DAEmut group sowed poorer survival than the α_DAEwt group, now supported with better number (n=35 vs 8, respectively). This supports the general idea that expressing more of the wild-type allele is better than expressing more of the mutated allele.

CHAPTER 6
CONCLUSIONS AND FUTURE PERSPECTIVES

6 Conclusions and Future Perspectives

In conclusion, in my study I show for the first time that there is a positive selection of cis-regulation of expression of mutated allele, both in *PIK3CA's* and *TP53's* mutated tumours. I also show an association between MADE of *PIK3CA's* mutant allele and the receptor status of breast cancer, especially PR status. There was also association between ER and PR status and MADE of *TP53's* mutant allele, but it was only observed in TCGA set. Furthermore, I show its association with overall and disease specific survivals of breast cancer patients, with γ ratios MADE defining an aggressive subset of *PIK3CA's* mutated tumours.

Since, this study does not unveil the mechanisms by which cis-regulation is acting, there is a need to study if the differential allelic expression observed here is due to epigenetic, germline or somatic alterations on cis-regulatory elements.

Furthermore, I focused on somatic mutations in coding regions.So the impact of mutations in non-coding regions was not evaluated. As previous studies have proposed that a substantial number of genes may have allele-specific expression with cis-regulatory variation, and that mutations in cis-regulatory regions can alter gene expression, whole-genome scale allelic imbalance analysis, including noncoding regions, will be required to elucidate the allele-specific expression of mutations beyond the coding regions.

Also, this study used bulk tumour RNA and DNA sequencing information, and since tumours are heterogeneous, having different clones with different mutations and gene expressions, it should be interesting to sample the different clones or assess single cell mutations expression, in order to discern whether the allelic imbalance observed in bulk tumour analysis is real or just the product of the combination of different mutant allelic expression levels of different groups of cells.

One of the greatest challenges in breast cancer treatment now a days is to determine which patients will benefit from more aggressive treatment, due to having a worst prognosis. Although substantial progress has been made in the past two decades, we still must find ways of identifying new biomarkers and ways of subdividing patients into different prognostic and response to treatment subgroups. Thus, the work presented here should be followed by the analysis of all mutations' allelic expression together, in an attempt to improve the existing molecular signatures, such as Integrative Clusters and PAM50 intrinsic subtypes.

The findings presented here support the idea that allelic imbalance generated by cis-regulation, besides that generated by copy number changes, has an essential role in modulating the effect of somatic mutations in tumours, and playing a part in oncogenesis. Thus, expression status of mutations should be taken into consideration to improve patient management in clinic, and general prognosis for such a common and deadly disease.

CHAPTER 7
REFERENCES

7 References

Adoue, V. *et al.* (2014) 'Allelic expression mapping across cellular lineages to establish impact of non-coding SNPs.', *Molecular systems biology*, 10, p. 754. doi: 10.15252/msb.20145114.

Alberts, B. *et al.* (2014) *Molecular Biology of the Cell 6e*, *Garland Science*. doi: 10.1002/1521-3773(20010316)40:6<9823::AID-ANIE9823>3.3.CO;2-C.

Aleskandarany, M. A. *et al.* (2010) 'PIK3CA expression in invasive breast cancer: A biomarker of poor prognosis', *Breast Cancer Research and Treatment*, 122(1), pp. 45–53. doi: 10.1007/s10549-009-0508-9.

Amin, M. *et al.* (2017) *AJCC Cancer Staging Manual*. 8th Editio. Springer International Publishing: American Joint Commission on Cancer.

Bachelot, T. *et al.* (2011) 'Mutational characterization of individual breast tumors: TP53 and PI3K pathway genes are frequently and distinctively mutated in different subtypes', *Breast Cancer Research and Treatment*, 132(1), pp. 29–39. doi: 10.1007/s10549-011-1518-y.

Batcha, A. M. N. *et al.* (2019) 'Allelic Imbalance of Recurrently Mutated Genes in Acute Myeloid Leukaemia', *Scientific Reports*. Springer US, 9(1), pp. 1–11. doi: 10.1038/s41598-019-48167-4.

Benitez, J. A., Cheng, S. and Deng, Q. (2017) 'Revealing allele-specific gene expression by single-cell transcriptomics', *International Journal of Biochemistry and Cell Biology*. Elsevier, 90(March), pp. 155–160. doi: 10.1016/j.biocel.2017.05.029.

Bewick, V., Cheek, L. and Ball, J. (2004) 'Statistics review 12: Survival analysis', *Critical Care*, 8(5), pp. 389–394. doi: 10.1186/cc2955.

Bhat-Nakshatri, P. *et al.* (2016) 'Molecular Insights of Pathways Resulting from Two Common PIK3CA Mutations in Breast Cancer', *Cancer Research*, 76(13), pp. 3989–4001. doi: 10.1158/0008-5472.CAN-15-3174.

Bielski, C. M. *et al.* (2018) 'Widespread Selection for Oncogenic Mutant Allele Imbalance in Cancer', *Cancer cell*. Elsevier Inc., 34(5), pp. 852–862. doi: 10.1016/j.ccell.2018.10.003.

Bland, K. and Copeland, E. (2009) *The Breast: Comprehensive Management of Benign and Malignant Diseases*. 4th edn.

Borresen-Dale, A.-L. (2003) 'TP53 and Breast Cancer', *HUMAN MUTATION*, 21, pp. 292–300. doi: 10.1002/humu.10174.

Bradburn, M. J. *et al.* (2003a) 'Survival Analysis Part II: Multivariate data analysis- An introduction to

concepts and methods', *British Journal of Cancer*, 89(3), pp. 431–436. doi: 10.1038/sj.bjc.6601119.

Bradburn, M. J. *et al.* (2003b) 'Survival Analysis Part III: Multivariate data analysis - Choosing a model and assessing its adequacy and fit', *British Journal of Cancer*, 89(4), pp. 605–611. doi: 10.1038/sj.bjc.6601120.

Bray, F. *et al.* (2018) 'Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries.', *CA: A Journal for Clinicians*. doi: 10.3322/caac.21492.

Bush, W. S. and Moore, J. H. (2012) 'Chapter 11: Genome-Wide Association Studies', *PLoS Computational Biology*, 8(12). doi: 10.1371/journal.pcbi.1002822.

Castel, S. E. *et al.* (2018) 'Modified penetrance of coding variants by cis-regulatory variation contributes to disease risk', *Nature Genetics*, 50(9), pp. 1327–1334. doi: 10.1038/s41588-018-0192-y.

Castle, J. C. *et al.* (2014) 'Mutated tumor alleles are expressed according to their DNA frequency', *Scientific Reports*, 4, p. 4743. doi: 10.1038/srep04743.

Cerami, E. *et al.* (2012) 'The cBio Cancer Genomics Portal : An Open Platform for Exploring Multidimensional Cancer Genomics Data', *Cancer Discovery*, 2(5), pp. 401–4. doi: 10.1158/2159-8290.CD-12-0095.

Chakravarty, D. *et al.* (2017) 'OncoKB: a precision oncology knowledge base', *JCO Precision OncologyPrecision Oncology*, 1, pp. 1–16. doi: 10.1200/PO.17.00011.

Chess, A. (2012) 'Mechanisms and consequences of widespread random monoallelic expression', *Nature Reviews Genetics*. Nature Publishing Group, 13(6), pp. 421–428. doi: 10.1038/nrg3239.

Cheung, V. G. *et al.* (2008) 'Monozygotic Twins Reveal Germline Contribution to Allelic Expression Differences', *American Journal of Human Genetics*, 82(6), pp. 1357–1360. doi: 10.1016/j.ajhg.2008.05.003.

Cheung, V. G. and Spielman, R. S. (2009) 'Genetics of human gene expression: Mapping DNA variants that influence gene expression', *Nature Reviews Genetics*, 10(9), pp. 595–604. doi: 10.1038/nrg2630.

Chia, S. *et al.* (2008) 'Human epidermal growth factor receptor 2 overexpression as a prognostic factor in a large tissue microarray series of node-negative breast cancers', *Journal of Clinical Oncology*, 26(35), pp. 5697–5704. doi: 10.1200/JCO.2007.15.8659.

Chiosea, S. I. *et al.* (2011) 'KRAS mutant allele-specific imbalance in lung adenocarcinoma', *Modern Pathology*. Nature Publishing Group, 24(12), pp. 1571–1577. doi: 10.1038/modpathol.2011.109.

Ciriello, G. *et al.* (2013) 'The molecular diversity of Luminal A breast tumors', *Breast Cancer Research and Treatment*, 141(3), pp. 409–420. doi: 10.1007/s10549-013-2699-3.

Clark, T. G. *et al.* (2003) 'Survival Analysis Part I: Basic concepts and first analyses', *British Journal of Cancer*, 89(2), pp. 232–238. doi: 10.1038/sj.bjc.6601118.

Cossu-Rocca, P. *et al.* (2015) 'Analysis of PIK3CA mutations and activation pathways in triple negative breast cancer', *PLoS ONE*, 10(11), pp. 1–14. doi: 10.1371/journal.pone.0141763.

Cox, D. G. *et al.* (2011) 'Common variants of the BRCA1 wild-type allele modify the risk of breast cancer in BRCA1 mutation carriers', *Human Molecular Genetics*, 20(23), pp. 4732–4747. doi: 10.1093/hmg/ddr388.

Curtis, C. *et al.* (2012) 'The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups', *Nature*, 486(7403), pp. 346–352. doi: 10.1038/nature10983.

Dudley, W. N., Wickham, R. and Coombs, N. (2016) 'An Introduction to Survival Statistics: Kaplan-Meier Analysis', *Journal of the Advanced Practitioner in Oncology*, 7(1), pp. 91–100. doi: 10.6004/jadpro.2016.7.1.8.

Duffy, M. J., Synnott, N. C. and Crown, J. (2018) 'Mutant p53 in breast cancer : potential as a therapeutic target and biomarker', *Breast Cancer Research and Treatment*. Springer US, (0123456789). doi: 10.1007/s10549-018-4753-7.

Dunnwald, L. K., Rossing, M. A. and Li, C. I. (2007) 'Hormone receptor status, tumor characteristics, and prognosis: A prospective cohort of breast cancer patients', *Breast Cancer Research*, 9(1), pp. 1–10. doi: 10.1186/bcr1639.

Eccles, S. A. *et al.* (2013) 'Critical research gaps and translational priorities for the successful prevention and treatment of breast cancer', *Breast Cancer Research*, 15(5), pp. 1–37. doi: 10.1186/bcr3493.

Eroglu, Z., Tagawa, T. and Somlo, G. (2014) 'Human Epidermal Growth Factor Receptor Family-Targeted Therapies in the Treatment of HER2-Overexpressing Breast Cancer', *The Oncologist*, 19(2), pp. 135–150. doi: 10.1634/theoncologist.2013-0283.

Forjaz de Lacerda, G. *et al.* (2018) 'Breast cancer in Portugal : Temporal trends and age-specific incidence by geographic regions', *Cancer Epidemiology*, 54, pp. 12–18. doi: 10.1016/j.canep.2018.03.003.

Fragomeni, S. M., Sciallis, A. and Jeruss, J. S. (2018) 'Molecular Subtypes and Local-Regional Control of Breast Cancer', *Surgical Oncology Clinics of North America*, 27(1), pp. 95–120. doi: 10.1016/j.soc.2017.08.005.

Gao, J. *et al.* (2013) 'Integrative Analysis of Complex Cancer Genomics and Clinical Profiles Using the cBioPortal', *Science Signaling*, 6(269), p. pl1. doi: 10.1126/scisignal.2004088.

George, B., Seals, S. and Aban, I. (2014) 'Survival analysis and regression models'. doi:

10.1007/s12350-014-9908-2.

Geroge, F. (2012) ‘Causas de Morte em Portugal e Desafios na Prevenção’, *Acta Médica Portuguesa*, 25(2), pp. 61–63.

Gonzalez-Zulueta, M. *et al.* (1995) ‘Methylation of the 5’ CpG island of the p16/CDKN2 tumor suppressor gene in normal and transformed human tissues correlates with gene silencing’, *Cancer Research*, 55(20), pp. 4531–4535. Available at: http://www.embase.com/search/results?subaction=viewrecord&from=export&id=L25306985.

Hanahan, D. and Weinberg, R. A. (2000) ‘The hallmarks of cancer.’, *Cell*, 100(1), pp. 57–70. doi: 10.1007/s00262-010-0968-0.

Hanahan, D. and Weinberg, R. A. (2011) ‘Hallmarks of Cancer: The Next Generation’, *Cell*, 144, pp. 646–674. doi: 10.1016/j.cell.2011.02.013.

Hartman, D. J. *et al.* (2012) ‘Mutant allele-specific imbalance modulates prognostic impact of KRAS mutations in colorectal adenocarcinoma and is associated with worse overall survival’, *International Journal of Cancer*, 131(8), pp. 1810–1817. doi: 10.1002/ijc.27461.

Kalinsky, K. *et al.* (2009) ‘PIK3CA mutation associates with improved outcome in breast cancer’, *Clinical Cancer Research*, 15(16), pp. 5049–5059. doi: 10.1158/1078-0432.CCR-09-0632.

Kasper, D. L. *et al.* (2015) *Harrison’s principles of internal medicine*. 19th Editi. New York: McGraw Hill Education.

Kassambara, A. and Kosinski, M. (2018) ‘survminer: Drawing Survival Curves using “ggplot2”. R package version 0.4.3.’

Kastenhuber, E. R. and Lowe, S. W. (2017) ‘Putting p53 in Context’, *Cell*. Elsevier Inc., 170(6), pp. 1062–1078. doi: 10.1016/j.cell.2017.08.028.

Keegan, N. M. *et al.* (2018) ‘PI3K inhibition to overcome endocrine resistance in breast cancer’, *Expert Opinion on Investigational Drugs*. Taylor & Francis, 27(1), pp. 1–15. doi: 10.1080/13543784.2018.1417384.

Kim, C. *et al.* (2004) ‘A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer.’, *The New England journal of medicine*, 351(27), pp. 2817–26. doi: 10.1056/NEJMoa041588.

Koboldt, D., Fulton, R. and McLellan, M. (2012) ‘Comprehensive molecular portraits of human breast tumours’, *Nature*, 490(7418), pp. 61–70. doi: 10.1038/nature11412.

Krasinskas, A. M. *et al.* (2013) ‘KRAS mutant allele-specific imbalance is associated with worse prognosis in pancreatic cancer and progression to undifferentiated carcinoma of the pancreas’, *Modern*

Pathology. Nature Publishing Group, 26(10), pp. 1346–1354. doi: 10.1038/modpathol.2013.71.

Leevers, S. J., Vanhaesebroeck, B. and Waterfield, M. D. (1999) 'Signalling through phosphoinositide 3-kinases: The lipids take centre stage', *Current Opinion in Cell Biology*, 11(2), pp. 219–225. doi: 10.1016/S0955-0674(99)80029-5.

Di Leo, A. *et al.* (2015) 'New approaches for improving outcomes in breast cancer in Europe', *Breast*, 24(4), pp. 321–330. doi: 10.1016/j.breast.2015.03.001.

Li, H. *et al.* (2015) 'Statistical inference methods for two crossing survival curves: A comparison of methods', *PLoS ONE*, 10(1), pp. 1–18. doi: 10.1371/journal.pone.0116774.

Li, Q. *et al.* (2013) 'Associations between single-nucleotide polymorphisms in the PI3K-PTEN-AKT-mTOR pathway and increased risk of brain metastasis in patients with non-small cell lung cancer', *Clinical Cancer Research*, 19(22), pp. 6252–6260. doi: 10.1158/1078-0432.CCR-13-1093.

Lim, B. E., Filho, O. M. and Winer, E. P. (2012) 'The Natural History of Hormone Receptor – Positive Breast Cancer', *Oncology*, 26(8), pp. 688–94.

Liu, J. *et al.* (2018) 'An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics', *Cell*, 173(2), pp. 400-416.e11. doi: 10.1016/j.cell.2018.02.052.

Liu, J., Zhang, C. and Feng, Z. (2013) 'Tumor suppressor p53 and its gain-of-function mutants in cancer The p53 Signaling Pathway Introduction Tumor Suppressive Functions of p53', *Acta biochimica et biophysica …*, 46(3), pp. 1–10. doi: 10.1093/abbs/gmt144.Review.

Liu, R. *et al.* (2012) 'Allele-specific expression analysis methods for high-density SNP microarray data', 28(8), pp. 1102–1108. doi: 10.1093/bioinformatics/bts089.

Lux, M. P. *et al.* (2016) 'The PI3K Pathway : Background and Treatment Approaches', pp. 398–404. doi: 10.1159/000453133.

Maia, A.-T. *et al.* (2012) 'Effects of BRCA2 cis-regulation in normal breast and cancer risk amongst BRCA2 mutation carriers', pp. 1–15.

Maia, A. *et al.* (2009) 'Extent of differential allelic expression of candidate breast cancer genes is similar in blood and breast', *Breast Cancer Research*, 11(6), pp. 1–10. doi: 10.1186/bcr2458.

Makki, J. (2015) 'Diversity of breast carcinoma: Histological subtypes and clinical relevance', *Clinical Medicine Insights: Pathology*, 8(1), pp. 23–31. doi: 10.4137/CPath.s31563.

Malapelle, U. *et al.* (2015) 'KRAS Mutant Allele-Specific Imbalance (MASI) assessment in routine samples of patients with metastatic colorectal cancer', *Journal of Clinical Pathology*, 68(4), pp. 265–269. doi: 10.1136/jclinpath-2014-202761.

McDonald, J. H. (2014) *Handbook of Biological Statistics*. 3rd edn. Sparky House Publishing.

Melton, C. *et al.* (2015) 'Recurrent somatic mutations in regulatory regions of human cancer genomes', *Nature Genetics*, 47(7), pp. 710–716. doi: 10.1038/ng.3332.

Mitsudomi, T. *et al.* (1991) 'ras gene mutations in non-small cell lung cancers are associated with shortened survival irrespective of treatment intent', *Cancer Research*, 51, pp. 4999–5002.

Moja, L. *et al.* (2012) 'Trastuzumab containing regimens for early breast cancer (Review) Summary of findings for main comparison', (4). doi: 10.1002/14651858.CD006243.pub2/abstract.

Morgese, F. *et al.* (2017) 'Impact of phosphoinositide-3-kinase and vitamin D3 nuclear receptor single-nucleotide polymorphisms on the outcome of malignant melanoma patients', *Oncotarget*, 8(44), pp. 75914–75923. doi: 10.18632/oncotarget.18304.

Nagaraj, G. and Ma, C. (2015) 'Revisiting the estrogen receptor pathway and its role in endocrine therapy for postmenopausal women with estrogen receptor-positive metastatic breast cancer', *Breast Cancer Research and Treatment*, 150(2), pp. 231–242. doi: 10.1007/s10549-015-3316-4.

NHS Cancer Screening Programmes and The Royal College of Pathologists (2005) *Pathology Reporting of Breast Disease: A Joint Document Incorporating the Third Edition of the NHS Breast Screening Programme's Guidelines for Pathology Reporting in Breast Cancer Screening and the Second Edition of The Royal College of Pathologists' Mini*. NHS Cancer Screening Programmes jointly with The Royal College of Pathologists.

Nicolini, A., Ferrari, P. and Duffy, M. J. (2018) 'Prognostic and predictive biomarkers in breast cancer: Past, present and future', *Seminars in Cancer Biology*. Elsevier Ltd, 52, pp. 56–73. doi: 10.1016/j.semcancer.2017.08.010.

Nik-zainal, S. *et al.* (2016) 'Landscape of somatic mutations in 560 breast cancer whole-genome sequences', *Nature*. Nature Publishing Group, 534(7605), pp. 47–54. doi: 10.1038/nature17676.

Norum, J. H., Andersen, K. and Sørlie, T. (2014) 'Lessons learned from the intrinsic subtypes of breast cancer in the quest for precision therapy', *British Journal of Surgery*, 101(8), pp. 925–938. doi: 10.1002/bjs.9562.

Nothnagel, M. *et al.* (2011) 'Statistical inference of allelic imbalance from transcriptome data', *Human Mutation*, 32(1), pp. 98–106. doi: 10.1002/humu.21396.

Oakley, G. J. and Chiosea, S. I. (2011) 'Higher dosage of the epidermal growth factor receptor mutant allele in lung adenocarcinoma correlates with younger age, stage IV at presentation, and poorer survival', *Journal of Thoracic Oncology*. International Association for the Study of Lung Cancer, 6(8), pp. 1407–1412. doi: 10.1097/JTO.0b013e31821d41af.

Pang, B. *et al.* (2014) 'Prognostic role of PIK3CA mutations and their association with hormone receptor expression in breast cancer: a meta-analysis', *Scientific Reports*, 4(1), p. 6255. doi: 10.1038/srep06255.

Parker, J. S. *et al.* (2009) 'Supervised risk predictor of breast cancer based on intrinsic subtypes', *Journal of Clinical Oncology*, 27(8), pp. 1160–1167. doi: 10.1200/JCO.2008.18.1370.

Pastinen, T. (2010) 'Genome-wide allele-specific analysis: Insights into regulatory variation', *Nature Reviews Genetics*. Nature Publishing Group, 11(8), pp. 533–538. doi: 10.1038/nrg2815.

Pastinen, T., Ge, B. and Hudson, T. J. (2006) 'Influence of human genome polymorphism on gene expression.', *Human molecular genetics*, 15 Spec No(1), pp. 9–16. doi: 10.1093/hmg/ddl044.

Pastinen, T. and Hudson, T. J. (2004) 'Cis-acting regulatory variation in the human genome', *Science*, 306(5696), pp. 647–650. doi: 10.1126/science.1101659.

Peduzzi, P. *et al.* (1995) 'Importance of events per independent variable in proportional hazards regression analysis II. Accuracy and precision of regression estimates', *Journal of Clinical Epidemiology*, 48(12), pp. 1503–1510. doi: 10.1016/0895-4356(95)00048-8.

Pereira, B. *et al.* (2016) 'The somatic mutation profiles of 2,433 breast cancers refines their genomic and transcriptomic landscapes', *Nature Communications*, 7(May). doi: 10.1038/ncomms11479.

Perou, C. M. C. M. *et al.* (2000) 'Molecular portraits of human breast tumours', *Nature*, 406(6797), pp. 747–752. doi: 10.1038/35021093.

Prat, A. and Perou, C. M. (2011) 'Deconstructing the molecular portraits of breast cancer', *Molecular Oncology*. Elsevier B.V, 5(1), pp. 5–23. doi: 10.1016/j.molonc.2010.11.003.

Press, D. (2015) 'PI3K mutations in breast cancer : prognostic and therapeutic implications', pp. 111–123.

Provenzano, E., Ulaner, G. A. and Chin, S. F. (2018) 'Molecular Classification of Breast Cancer', *PET Clinics*. Elsevier Inc, 13(3), pp. 325–338. doi: 10.1016/j.cpet.2018.02.004.

Pu, X. *et al.* (2011) 'PI3K/PTEN/AKT/mTOR pathway genetic variation predicts toxicity and distant progression in lung cancer patients receiving platinum-based chemotherapy', *Lung Cancer*. Elsevier Ireland Ltd, 71(1), pp. 82–88. doi: 10.1016/j.lungcan.2010.04.008.

Qiu, P. and Sheng, J. (2008) 'A two-stage procedure for comparing hazard rate functions', *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 70(1), pp. 191–208. doi: 10.1111/j.1467-9868.2007.00622.x.

Rakha, E. A. *et al.* (2010) 'Breast cancer prognostic classification in the molecular era: The role of histological grade', *Breast Cancer Research*, 12(4). doi: 10.1186/bcr2607.

Reisenbichler, E. S. *et al.* (2017) 'Is tumor cellularity in primary invasive breast carcinoma of prognostic significance?', *Virchows Archiv*. Virchows Archiv, 470(6), pp. 611–617. doi: 10.1007/s00428-017-2120-4.

Rhee, J. *et al.* (2017) 'Allelic imbalance of somatic mutations in cancer genomes and transcriptomes', *Scientific Reports*. Springer US, 7, p. 1653. doi: 10.1038/s41598-017-01966-z.

Rheinbay, E. *et al.* (2017) 'Recurrent and functional regulatory mutations in breast cancer', *Nature*. Nature Publishing Group, 547(7661), pp. 55–60. doi: 10.1038/nature22992.

Rivandi, M., Martens, J. W. M. and Hollestelle, A. (2018) 'Elucidating the Underlying Functional Mechanisms of Breast Cancer Susceptibility Through Post-GWAS Analyses', *Frontiers in Genetics*, 9(August). doi: 10.3389/fgene.2018.00280.

Rueda, O. M. *et al.* (2019) 'Dynamics of breast-cancer relapse reveal late-recurring ER-positive genomic subgroups', *Nature*. Springer US, 567, pp. 399–404. doi: 10.1038/s41586-019-1007-8.

Russnes, H. G. *et al.* (2017) 'Breast Cancer Molecular Stratification: From Intrinsic Subtypes to Integrative Clusters', *American Journal of Pathology*, 187(10), pp. 2152–2162. doi: 10.1016/j.ajpath.2017.04.022.

Sabine, V. S. *et al.* (2014) 'Mutational analysis of PI3K/AKT signaling pathway in tamoxifen exemestane adjuvant multinational pathology study', *Journal of Clinical Oncology*, 32(27), pp. 2951–2958. doi: 10.1200/JCO.2013.53.8272.

Sansal, I. and Sellers, W. R. (2004) 'The biology and clinical relevance of the PTEN tumor suppressor pathway', *Journal of Clinical Oncology*, 22(14), pp. 2954–2963. doi: 10.1200/JCO.2004.02.141.

Santarpia, L. *et al.* (2016) 'Deciphering and Targeting Oncogenic Mutations and Pathways in Breast Cancer', *The Oncologist*, 21(9), pp. 1063–1078. doi: 10.1634/theoncologist.2015-0369.

Schaid, D. J., Chen, W. and Larson, N. B. (2018) 'From genome-wide associations to candidate causal variants by statistical fine-mapping', *Nature Reviews Genetics*. Springer US, 19(8), pp. 491–504. doi: 10.1038/s41576-018-0016-z.

Schnitt, S. J. (2010) 'Classification and prognosis of invasive breast cancer: From morphology to molecular taxonomy', *Modern Pathology*. Nature Publishing Group, 23(S2), pp. 60–64. doi: 10.1038/modpathol.2010.33.

Schober, P. and Vetter, T. R. (2018) 'Survival analysis and interpretation of time-to-event data: The tortoise and the hare', *Anesthesia and Analgesia*, 127(3), pp. 792–798. doi: 10.1213/ANE.0000000000003653.

Sigal, A. and Rotter, V. (2000) 'Oncogenic mutations of the p53 tumor suppressor: the demons of the

guardian of the genome.', *Cancer research*, 60(24), pp. 6788–93. Available at: http://www.ncbi.nlm.nih.gov/pubmed/11156366.

Soh, J. *et al.* (2009) 'Oncogene mutations, copy number gains and mutant allele specific imbalance (MASI) frequently occur together in tumor cells', *PLoS ONE*, 4(10), p. e7464. doi: 10.1371/journal.pone.0007464.

Sorlie, T. *et al.* (2001) 'Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications', *Proceedings of the National Academy of Sciences of the United States of America*, 98(19), pp. 10869–10874. doi: 10.1073/pnas.191367098.

Soussi, T. and Wiman, K. G. (2015) 'TP53: An oncogene in disguise', *Cell Death and Differentiation*. Nature Publishing Group, 22(8), pp. 1239–1249. doi: 10.1038/cdd.2015.53.

Stagni, C. *et al.* (2018) 'BRAF gene copy number and mutant allele frequency correlate with time to progression in metastatic melanoma patients treated with MAPK inhibitors.', *Molecular Cancer Therapeutics*, 17(June), p. molcanther.1124.2017. doi: 10.1158/1535-7163.MCT-17-1124.

Stephens, P. J. *et al.* (2012) 'The landscape of cancer genes and mutational processes in breast cancer', *Nature*, 486(7403), pp. 400–404. doi: 10.1038/nature11017.

Sun, W. *et al.* (2016) 'Nomograms to estimate long-term overall survival and breast cancer-specific survival of patients with luminal breast cancer.', *Oncotarget*, 7(15), pp. 20496–506. doi: 10.18632/oncotarget.7975.

Tao, Z. Q. *et al.* (2015) 'Breast Cancer: Epidemiology and Etiology', *Cell Biochemistry and Biophysics*. Springer US, 72(2), pp. 333–338. doi: 10.1007/s12013-014-0459-6.

The Cancer Genome Atas Research (2013) 'Genomic and Epigenomic Landscapes of Adult De Novo Acute Myeloid Leukemia', *Tne New England Journal of Medicine*, 368(22), pp. 2059–74. doi: 10.1056/NEJMoa1301689.

The International Hapmap Consortium (2003) 'The International HapMap Project', *Nature*, 426(6968), pp. 789–796. doi: 10.1038/nature02168.

Valencia, O. M. *et al.* (2017) 'The role of genetic testing in patients with breast cancer a review', *JAMA Surgery*, 152(6), pp. 589–594. doi: 10.1001/jamasurg.2017.0552.

Végran, F. *et al.* (2013) 'Only Missense Mutations Affecting the DNA Binding Domain of P53 Influence Outcomes in Patients with Breast Carcinoma', *PLoS ONE*, 8(1), p. e55103. doi: 10.1371/journal.pone.0055103.

Véron, A., Blein, S. and Cox, D. G. (2014) 'Genome-wide association studies and the clinic: A focus on breast cancer', *Biomarkers in Medicine*, 8(2), pp. 287–296. doi: 10.2217/bmm.13.121.

Wang, G. *et al.* (2009) 'Knockin of mutant PIK3CA activates multiple oncogenic pathways', *Proceedings of the National Academy of Sciences*, 106(8), pp. 2835–2840. doi: 10.1073/pnas.0813351106.

Wang, L. E. *et al.* (2012) 'Roles of genetic variants in the PI3K and RAS/RAF pathways in susceptibility to endometrial cancer and clinical outcomes', *Journal of Cancer Research and Clinical Oncology*, 138(3), pp. 377–385. doi: 10.1007/s00432-011-1103-0.

Wang, W. *et al.* (2017) 'The impact of heterogeneity in phosphoinositide 3-kinase pathway in human cancer and possible therapeutic treatments', *Seminars in Cell and Developmental Biology*. Elsevier Ltd, 64, pp. 116–124. doi: 10.1016/j.semcdb.2016.08.024.

Weinberg, R. A. (2013) *The Biology of Cancer_2nd edition*, *Journal of Chemical Information and Modeling*. doi: 10.1017/CBO9781107415324.004.

Wickham, H. (2016) *ggplot2: Elegant Graphics for Data Analysis*. 1st Editio. Springer-Verlag New York.

Xiao, R. and Scott, L. J. (2011) 'Detection of cis-acting regulatory SNPs using allelic expression data', *Genetic Epidemiology*, 35(6), pp. 515–525. doi: 10.1002/gepi.20601.

Yip, C.-H. and Rhodes, A. (2014) 'Estrogen and progesterone receptors in breast cancer', *Future Oncology*, 10(14), pp. 2293–2301.

Youn, S. *et al.* (2018) 'Differences in prognosis and efficacy of chemotherapy by p53 expression in triple-negative breast cancer', *Breast Cancer Research and Treatment*. Springer US, 0(0), p. 0. doi: 10.1007/s10549-018-4928-2.

Yuan, T. L. and Cantley, L. C. (2008) 'PI3K pathway alterations in cancer: Variations on a theme', *Oncogene*, 27(41), pp. 5497–5510. doi: 10.1038/onc.2008.245.

Zardavas, D., Phillips, W. A. and Loi, S. (2014) 'PIK3CA mutations in breast cancer : reconciling findings from preclinical and clinical data', *Breast cancer research : BCR*, 2, pp. 1–10.

ANNEXES

Annex A

Table A.1: Summary of α, β and γ ratios of all genes with more than 2 samples with DNAseq and RNAseq data from METABRIC dataset. SD: Standard deviation; Min: minimum value of the ratio; Max: maximum value of the ratio.

		α				β				γ						
Hugo Symbol	**count**	**Mean**	**SD**	**Min**	**Max**	**Mean**	**SD**	**Min**	**Max**	**Mean**	**SD**	**Min**	**Max**	**MADE_wt**	**mut=wt**	**MADE_mut**
PIK3CA	115	-0.7759	1.2022	-2.5236	5.7549	-1.2788	1.1938	-3.6666	5.1714	0.5029	0.8071	-1.2805	3.6969	10	57	48
TP53	69	1.2602	1.7029	-2.3219	4.9542	-0.4945	1.3605	-3.3735	2.9583	1.8193	1.5881	-1.8536	5.4589	7	8	54
AKT1	13	0.8933	1.6801	-1.8931	3.6856	-0.2729	1.4611	-2.8882	2.009	1.1661	1.1166	-0.4443	3.0987	0	5	8
GATA3	13	-0.298	1.8298	-2.0458	3.2159	-1.2017	1.0505	-3.0311	0.3304	0.9037	1.6175	-1.7898	3.7709	2	4	7
PTEN	9	-0.748	1.1714	-2.4053	0.8519	-0.8869	1.161	-2.6921	0.1213	0.139	0.9249	-1.1203	1.8387	2	4	3
ERBB2	8	0.3212	1.2845	-1.5937	2.8822	-0.61	1.3758	-3.061	0.9827	0.9312	1.056	-0.0983	2.9855	0	4	4
CDH1	6	4.286	3.3313	-2.175	6.864	-0.0331	0.9067	-1.7925	0.6392	4.319	3.6363	-2.688	7.032	1	0	5
MAP2K4	5	-0.2188	1.6449	-2.2955	2.2395	-2.245	1.243306	-3.373	-0.555	2.0259	1.12836	0.7372	3.5312	0	0	5
MYH9	4	-1.1973	1.2186	-2.8231	-0.189	-1.0646	1.4952	-3.0163	0.4125	-0.1326	0.3348	-0.6015	0.1931	1	3	0
FBXW7	3	1.9	2.1523	-0.585	3.143	0.8618	1.7285	-1.1341	1.8598	1.0385	0.4237	0.5492	1.2831	0	1	2
MAP3K1	3	-0.9795	0.6615	-1.7411	-0.5475	-1.5	0.2764	-1.793	-1.243	0.5204	0.8892	-0.498	1.1425	0	1	2
SF3B1	3	-0.7319	0.9832	-1.6245	0.3219	-1.8908	1.9677	-4.1593	-0.6453	1.1589	1.2907	-0.0253	2.5348	0	1	2
CHEK2	2	-0.432	0.7711	-0.9773	0.1132	-1.427	0.2996	-1.639	-1.216	0.9954	0.4714	0.6621	1.3287	0	0	2
ERBB3	2	-0.1678	0.4074	-0.4558	0.12029	-0.0432	0.0285	-0.0635	-0.0231	-0.1244	0.3788	-0.3923	0.1433	0	2	0

Table A.2: Summary of α, β and γ ratios of all genes with more than 10 samples with DNAseq and RNAseq data from TCGA dataset. SD: Standard deviation; Min: minimum value of the ratio; Max: maximum value of the ratio.

		α				β				γ						
Hugo Symbol	**count**	**Mean**	**SD**	**Min**	**Max**	**Mean**	**SD**	**Min**	**Max**	**Mean**	**SD**	**Min**	**Max**	**MADE_wt**	**mut=wt**	**MADE_mut**
PIK3CA	213	-0.813104513	1.477287017	-4.824428435	7.781359714	-1.072537033	1.314152777	-5.566054038	5.820875188	0.25943252	0.829110955	-2.101962411	3.068603814	30	115	68
TP53	150	1.595212167	2.244499217	-5.044394119	6.813781191	-0.251116187	1.477547629	-4.95419631	5.781359714	1.846328354	1.889267761	-3.924099886	5.076102979	20	10	120
GATA3	68	-0.681476508	1.513537366	-8.451211112	1.723188456	-1.649161188	0.923851504	-3.852442812	0.773724144	0.96768468	1.683012537	-7.866248611	5.575631267	7	12	49
MAP3K1	66	-1.454996888	2.109423696	-5.727920455	4.485426827	-1.978352453	1.433342193	-5.718818247	1.025995209	0.523355565	1.985780072	-4.86391691	3.887167504	11	17	38
NBPF10	30	-3.950969089	1.163006423	-7.055282436	-1.836501268	-5.060078597	2.674637909	-9.612868497	-2.273018494	1.109109508	2.759721333	-3.218640286	5.63879598	10	9	11
EPPK1	26	-3.467381615	1.454298003	-7.651051691	-0.235378063	-3.764133106	0.961164123	-5.21916852	-1.078609835	0.296751491	1.495765955	-4.20259119	2.219227921	4	9	13
PTTG1	26	-2.179033735	0.739405958	-3.442943496	-0.273284503	-4.820661083	1.459593947	-6.965784285	-0.476438044	2.641627347	1.501439219	-0.121015401	5.911336501	0	2	24
USP34	26	-3.120543495	1.332584799	-5.189824559	-0.27237227	-2.547104701	0.931946594	-4.285402219	-0.662304589	-0.573438794	0.922045057	-2.623477736	0.611434712	12	12	2
SON	23	-3.052473247	1.102633772	-4.59442283	-0.78379145	-5.808993475	2.251558428	-8.13442632	-1.129283017	2.756520228	1.665712605	-0.023567177	5.054447784	0	5	18
ARID1A	22	-1.834507171	1.232956452	-4.727920455	0	-2.597661565	1.971141638	-6.584962501	0.186413124	0.763154395	1.47256635	-2.034269902	3.326228232	6	7	9
CDH1	22	0.055154549	2.725504943	-4.087462841	5.988684687	-0.699931911	1.512479173	-2.795180208	3.169925001	0.75508646	2.249073126	-2.107640723	5.931765264	7	5	10
KIAA0100	22	-3.500719449	0.894882282	-4.768184325	-1.015596855	-5.780011304	1.108222923	-7	-2.715161224	2.279291855	0.985177012	0.191599268	4.364836686	0	1	21
ZNF814	22	-2.26228234	0.623019843	-3.440572591	-1.337034987	-7.934592056	0.803444665	-10.00702727	-6.906890596	5.672309716	0.991853561	4.027032959	7.099201678	0	0	22
RUNX1	21	-1.597163276	2.882389555	-4.89077093	7.30833903	-1.619391238	1.526301475	-6.906890596	0.910732662	0.022227962	3.284615986	-3.727956318	10.04818713	13	2	6
AHNAK	19	-2.477319207	1.791288899	-6.505811554	-0.024423474	-2.688451652	1.659300714	-6.311586151	-0.276709343	0.211132445	1.21051155	-2.837563135	2.126192931	3	8	8
EEF1D	19	-4.262471629	1.469748974	-5.591958611	-0.607361886	-5.690638497	2.093482408	-7.982993575	-0.980371193	1.428166869	1.049575145	-0.279622898	3.647389503	0	4	15
MUC2	18	-4.945293023	1.065773945	-6.304389352	-1.870020794	-6.336071224	1.834017905	-9.511752654	-2.494764692	1.390778201	2.191628316	-3.809016056	5.759335826	4	1	13
NCOR1	18	-0.908813561	2.14559821	-6.330916878	3.542527234	-0.788691903	1.59049935	-4.226068079	2.232660757	-0.120121658	1.604363222	-2.862141089	2.893008887	6	5	7
USP9X	18	-2.154689044	1.823026734	-4.874469118	1.710493383	-2.048752841	1.570685362	-5.569855608	0.061400545	-0.105936203	1.859703904	-4.319408103	2.991652657	6	7	5
ERBB2	17	-1.771342196	2.081103453	-5.100038852	1.403124178	-2.528522216	2.039187675	-6.475733431	0.584962501	0.757180021	0.942733026	-0.571086777	2.832890014	0	9	8
GOLGB1	17	-2.080890858	1.556896478	-4.153805336	1.067114196	-2.461611769	1.38679663	-5.882643049	-0.103093493	0.380720911	1.253169237	-1.80866985	2.582535763	4	5	8
SQSTM1	17	-5.260247583	0.96542164	-5.962173031	-2.905277745	-4.435258889	0.592982765	-5.169925001	-3.321928095	-0.824988694	1.159394046	-2.301858665	1.665132849	12	2	3
IRF2BP2	16	-4.76186169	0.251715189	-5.161887682	-4.044394119	-4.734541641	1.009101795	-6.700439718	-3.273018494	-0.027320049	1.136303854	-1.481947353	2.656045599	6	6	4
MACF1	16	-2.527192658	1.425011571	-6.122396631	0.078002512	-2.948217211	1.421516706	-4.938599455	-0.229481846	0.421024553	1.425736391	-1.819427754	3.028569152	4	6	6
MLL3	16	-1.3181537	1.489816187	-3.944858446	1.94753258	-1.584640693	0.754041105	-3.355480655	-0.555389385	0.266486994	1.277085584	-1.837894503	3.366033935	4	6	6
PTEN	16	-1.207637346	1.744166028	-5.988684687	1.365649472	-1.626106463	1.268864348	-3.866733469	0.736965594	0.418469117	1.348094351	-2.121951218	3.352301744	5	3	8
RBX1	16	-3.988149484	0.506905825	-5.028408415	-3.224630894	-4.387304241	1.221282216	-6.942514505	-2.700439718	0.399154757	1.295140889	-1.218707375	2.908501646	3	7	6
TRPS1	16	-3.322689066	1.778459391	-6.66106548	-0.437063806	-4.174332979	1.780948723	-7.554588852	-1.337034987	0.851643912	0.996338351	-0.947543299	3.181130456	2	3	11
ZNF28	16	-1.484185268	1.10699692	-3.95419631	0.160464672	-5.464283469	2.734545244	-8.761551232	0.401098308	3.9800982	2.547019654	-0.401098308	6.859062071	0	2	14
CTCF	15	-0.089966918	2.259382566	-4.48112669	2.429987841	-0.982972854	1.467660421	-3.832890014	1.299560282	0.893005936	1.785507721	-3.385969457	2.943055753	3	1	11
HECTD1	15	-1.2984829	1.967413175	-3.459431619	3.36923381	-1.833722375	1.991707023	-5.375039431	2.680119734	0.535239475	1.166504267	-0.94596016	2.539616924	3	6	6
MALAT1	15	-2.208541202	1.564115545	-4.273018494	0.222392421	-2.974922292	1.631633174	-5.209453366	-0.830074999	0.766381089	1.002937938	-0.222392421	3.919946748	0	9	6
SYNE2	15	-1.478754721	1.018797764	-2.979822118	0.862496476	-2.612343373	1.31974998	-5.209453366	-0.821662759	1.133588653	0.819481367	-0.623116517	2.316827223	1	3	11
TPR	15	-2.533667819	1.186836945	-4.581953751	-0.639824436	-3.09044546	1.362487513	-5.888743249	-1.070389328	0.556777641	0.809820387	-0.80407979	1.557923248	2	5	8
ARFGEF1	14	-2.59373438	0.962377714	-4.321928095	-1.074767768	-3.053483893	1.194260665	-5.169925001	-1.367371066	0.459749514	0.638991577	-1.110810357	1.7589919	1	9	4
BPNT1	14	-2.547133356	0.657837094	-3.807354922	-1.788495895	-5.736060221	1.140519084	-7.321928095	-2.775293713	3.188926865	1.247810097	0.981297174	5.02075856	0	0	14
CHD4	14	-2.085133671	2.274525734	-6.718818247	1.646363045	-2.171185382	1.589515522	-5.026800059	0.703282468	0.086051711	2.040417941	-4.514753498	2.509164071	2	6	6
FASN	14	-2.542960808	1.764604898	-5.11042399	-0.205318908	-3.379746229	1.620180758	-4.95419631	-0.231325546	0.836785421	1.213892465	-0.862496476	2.446221307	2	4	8

		α				β				γ						
Hugo Symbol	**count**	**Mean**	**SD**	**Min**	**Max**	**Mean**	**SD**	**Min**	**Max**	**Mean**	**SD**	**Min**	**Max**	**MADE_wt**	**mut=wt**	**MADE_mut**
SF3B1	14	-1.662295965	1.84027418	-5.044394119	1.902702799	-1.624409259	1.365032258	-4.352516415	0.292781749	-0.037886706	0.877717868	-1.844254505	1.609921049	4	8	2
UBR5	14	-1.890988605	1.626108974	-4.423681594	1.142019005	-2.174417625	1.756147579	-4.098733954	3.084064265	0.28342902	0.989854892	-1.94204526	1.788598363	2	7	5
ADNP	13	-2.297448679	0.808678846	-3.948113258	-0.584962501	-2.89312828	1.148339159	-5.163746427	-1.514573173	0.595679601	0.707560575	-0.630198351	1.550064293	2	3	8
CMPK1	13	-3.619330057	0.991415379	-5.22881869	-2.129283017	-4.656896951	1.632086287	-7.531381461	-2.736965594	1.037566894	2.089032712	-1.687060688	4.239305804	3	4	6
DDX5	13	-3.447473076	1.760852085	-5.880366607	-1.226526221	-2.703955586	1.560914336	-5.209453366	-0.944290567	-0.74351749	1.731658869	-3.359855787	1.294037055	5	5	3
MED1	13	-5.491115355	1.402922323	-8.082149041	-2.662965013	-5.438890497	1.172522278	-7.104598754	-3.483815777	-0.052224857	1.342414146	-2.490577173	1.869937495	4	4	5
METTL2A	13	-3.750106843	1.247162083	-7.189824559	-2.559427409	-6.797521413	1.590703125	-8.686500527	-2.636863603	3.04741457	2.40993637	-4.552960956	5.010935478	1	0	12
NBAS	13	-1.145933226	1.630073818	-3.584962501	2.980547637	-1.63339277	1.810273827	-3.722466024	3.140177658	0.487459544	0.981696033	-0.956931278	2.530514717	1	7	5
NUP160	13	-2.209911975	1.330612663	-3.882643049	-0.142444265	-2.818941303	1.603753221	-5.321928095	-0.299560282	0.609029328	0.804662645	-1.5898613	1.649502753	1	5	7
SPEN	13	-0.641331301	2.159203163	-3.95419631	1.700439718	-1.527567792	1.21811211	-4.216317907	0.273018494	0.886236491	1.604627812	-2.088809267	3.426264755	2	2	9
SUCO	13	-2.8105837	1.720086376	-5.914883386	-0.514573173	-3.543299949	2.132954922	-7.982993575	-1.105116706	0.732716248	1.132884877	-0.39132143	3.436505222	0	7	6
AKT1	12	1.191081094	1.684420134	-1.432111013	3.651363726	0.347867321	1.244055056	-1.353636955	2.862496476	0.843213774	1.000575578	-1.253253385	2.227930605	1	4	7
ARID4B	12	-2.915339082	1.585837144	-5.297680549	0.30256277	-3.355715399	1.725560747	-7.098032083	-0.368387406	0.440376317	1.393472243	-2.218383238	3.528176475	2	4	6
BIRC6	12	-1.660244719	1.440018661	-4.922832139	0.353636955	-2.327329379	1.434474456	-5.026800059	-0.206450877	0.66708466	0.667582635	-0.197510833	2.169925001	0	7	5
CAD	12	-2.041529231	1.27108093	-4.112700133	-0.215012891	-2.465196203	1.220021978	-4.273018494	-0.762960803	0.423666973	0.568539716	-0.598834675	1.575684687	1	8	3
CHD6	12	-1.892583985	1.264733579	-4.087462841	0.041820176	-2.075257917	1.095235527	-3.468148836	-0.365649472	0.182673932	0.752104501	-0.636434615	1.066221519	2	5	5
CIRBP	12	-1.309906941	0.733483627	-2.769387072	-0.560714954	-1.980547238	0.782425719	-3.029747343	-0.087462841	0.670640297	1.04195388	-1.676607983	1.621488377	1	3	8
DDB1	12	-1.615039488	2.113287237	-4.432959407	1.128007612	-2.29918719	1.808231605	-4.72631835	-0.625604485	0.684147702	0.612958827	0.148328773	1.89503044	0	6	6
FAM208B	12	-2.224602799	2.076354025	-4.096861539	3.513490746	-2.72506469	2.541624913	-8.011227255	2.243925583	0.500461891	1.647549079	-2.552541023	3.914365716	3	4	5
FLNB	12	-1.846490578	2.168733598	-6.242333497	2.592457037	-2.39346928	1.748420057	-6.048759312	0.185555653	0.546978702	1.461964649	-3.333641515	2.406901384	1	4	7
KIAA1467	12	-2.536000593	1.222875686	-4.466699619	-0.801454321	-4.622362212	1.453066734	-6.845490051	-1.5360529	2.086361618	1.447770875	0.126912112	4.652076697	0	2	10
MYH9	12	-2.059189621	1.50230926	-5.314257082	-0.614957241	-2.088191369	1.605543926	-6.22881869	-0.691877705	0.029001748	1.633692333	-4.392259594	2.069620096	2	5	5
ODF2	12	-1.817764297	1.332025256	-4.584962501	0	-2.438310363	1.442376199	-5.882643049	-0.688055994	0.620546066	1.078536794	-2.148445274	1.816553859	1	3	8
PIK3R1	12	-1.215567328	1.623277313	-4.361456459	1.812372997	-2.13824318	1.651397477	-5.330916878	-0.04580369	0.922675852	1.985419695	-0.919168925	5.830495499	1	7	4
UBA1	12	-2.019842121	1.730165076	-4.614295098	1.102759574	-2.338721773	1.756134489	-4.95419631	0.05246742	0.318879652	1.727962161	-2.798569004	3.102447269	4	1	7
WWP1	12	-3.904298171	1.074773809	-5.024585638	-1.954770746	-5.13638242	2.612554307	-10.30378075	-2.007819504	1.232084249	2.007301955	-1.519528055	5.344693941	1	6	5
APP	11	-3.472128188	1.466109309	-5.745954377	-1.598001796	-3.918485147	1.359192409	-6.224966365	-1.157541277	0.446356958	1.201882377	-0.954619399	2.675016699	2	7	2
ARHGAP35	11	-1.590446506	1.239266451	-3.966833136	-0.284567567	-2.161735984	1.45192407	-5.357552005	-0.768674454	0.571289478	0.789349964	-1.066474278	1.440572591	1	3	7
CCNI	11	-3.741351604	1.295288826	-5.69656324	-1.558908609	-7.587781037	0.910431571	-8.751544059	-5.857980995	3.846429433	1.233526539	1.812820097	5.798643396	0	0	11
COL12A1	11	-5.201538931	2.355670831	-9.894817763	-0.386217223	-2.281674206	1.694903211	-6.357552005	0.108252891	-2.919864724	2.764682585	-7.602036014	3.069307036	9	1	1
DSP	11	-2.139234823	2.315487914	-5.614709844	1.949959318	-1.563075352	1.077183781	-3.115477217	-0.222392421	-0.576159471	2.422668176	-4.94753258	2.409390936	4	3	4
EIF4A2	11	-2.113296477	2.948354121	-5.242856524	3.273018494	-2.929038522	2.216831005	-6.50779464	0.442004547	0.815742045	1.75241445	-1.962748605	4.980837743	2	3	6
ERBB3	11	-1.239585001	2.480624703	-5.874469118	2.874469118	-2.211354742	1.855042876	-5.382918725	1	0.971769741	1.644225866	-2.331205908	3.731803889	1	3	7
FRG1B	11	-2.764406462	0.778971597	-4.209453366	-1.632268215	-4.382647838	1.469727183	-5.857980995	-1.237039197	1.618241375	1.483331313	-0.610497593	3.749456538	1	2	8
GCC2	11	-1.965586786	1.553199585	-4.169925001	1	-4.215042216	3.172390868	-8.968666793	-0.299560282	2.249455429	2.700692734	-0.732777797	7.145544555	1	2	8
GPR160	11	-3.39960854	1.155418215	-4.332983283	-0.314873337	-4.931744936	1.525535046	-6.8008999	-1.169925001	1.532136396	1.092263998	-0.518087657	3.771152557	0	1	10
HUWE1	11	-1.146266149	3.170543897	-7.751544059	3.514573173	-2.336801603	1.146439608	-4.235216462	-1.038135129	1.190535454	2.95498292	-6.71340893	4.67211445	1	1	9
IL6ST	11	-2.173305031	2.924739819	-6.386242908	3.614709844	-2.690249914	2.978072378	-7.247927513	1.779231321	0.516944882	1.834033422	-4.126991205	2.465187891	1	6	4

		α				β				γ						
Hugo Symbol	count	Mean	SD	Min	Max	Mean	SD	Min	Max	Mean	SD	Min	Max	MADE_wt	mut=wt	MADE_mut
NOL6	11	-1.719577359	0.760287061	-2.672425342	0.169925001	-2.189356055	1.152380709	-4.554588852	-0.652076697	0.469778696	0.902189399	-0.888968688	1.969626351	1	5	5
PRRC2C	11	-1.465207217	1.408159812	-3.62058641	0.79970135	-1.93498774	1.680722343	-5.392317423	0.611929548	0.469780524	0.993256368	-1.277356652	1.906890596	2	5	4
RNF213	11	-1.782588412	1.840144253	-6.048032696	0.310340121	-2.126392657	1.754011261	-5.787902559	1.028014376	0.343804245	0.862133716	-1.026812473	1.811579508	2	5	4
SMEK1	11	-3.667639603	1.016737826	-4.733354341	-0.962197967	-5.647663027	1.702041337	-7.17990909	-1.289506617	1.980023424	1.03004792	0.154611946	3.432111013	0	2	9
SNRNP200	11	-1.812442463	1.393163515	-5.123382416	-0.459431619	-2.58852073	1.198301759	-3.95419631	-1.098180394	0.776078267	1.188625016	-1.663950797	2.504374579	1	3	7
SRRM2	11	-1.988495872	1.879296669	-4.169925001	1.416895445	-2.393572997	2.187077205	-5.672425342	1.206450877	0.405077124	0.808043727	-1.483686127	1.700994494	1	6	4
TBX3	11	-1.725466527	2.698140031	-6.007494537	2.5360529	-2.478675587	1.444593274	-4.887525271	-0.693896872	0.75320906	2.172433411	-3.598825154	3.36732187	3	2	6
TLN1	11	-2.423587989	2.211334924	-4.824428435	3.302119614	-2.115516585	2.092355775	-5.523561956	2.415037499	-0.308071404	1.756413431	-4.293913719	2.064130337	4	3	4
XPOT	11	-1.505172397	1.115316642	-3.689034955	0.712718048	-2.353709728	1.284215931	-4.413302448	0.566346823	0.848537331	0.405536836	0.146371225	1.300866479	0	2	9
ZNF217	11	-2.365163046	1.260271208	-4.235888264	-0.266280065	-3.114398802	2.343299678	-7.912889336	-0.462971976	0.749235756	1.944995213	-1.463333056	4.912889336	2	5	4
ASH1L	10	-2.81410864	1.617348603	-5.235216462	-0.637429921	-3.368807541	1.824234607	-5.54303182	-0.73039294	0.554698901	0.646039572	-0.471823525	1.773644748	0	4	6
ATP1B1	10	-2.381722393	1.55471709	-3.91753784	1.075288127	-3.788937038	2.215098767	-7.607330314	-0.054447784	1.407214645	1.690435679	-0.276029201	5.159871337	0	4	6
CUL7	10	-2.191158885	1.159731883	-3.8259706	-0.243925583	-3.789635998	2.039683825	-7.330916878	-1.14974712	1.598477113	1.359744383	-0.099535674	3.504946278	0	3	7
EIF3E	10	-3.943712571	1.506501454	-5.700439718	-1.829511336	-4.867286505	2.501273174	-9.108524457	-2.192645078	0.923573934	1.982203357	-1.612976877	3.879962268	2	5	3
FAM208A	10	-2.323128184	1.24100039	-4.129283017	0.097297201	-3.865423141	2.133318903	-6.599912842	-0.874469118	1.542294957	1.774564255	-0.894817763	3.237039197	1	3	6
GON4L	10	-1.81727202	1.291534531	-3.672425342	0.719892081	-2.513863634	1.472340028	-4.813781191	0.122484007	0.696591614	0.922796041	-0.774188058	1.796110165	1	2	7
IGSF8	10	-1.667889769	0.378196741	-2.072756342	-1.199308808	-2.62871322	1.340131379	-5.129283017	-1.61667136	0.960823451	1.24479871	-0.456084982	3.098256121	0	4	6
IKBKAP	10	-1.649648989	1.355726598	-3.765534746	0.090197809	-2.546560441	1.260611928	-4.668378509	-1.013546532	0.896911453	0.8394415	0.070389328	3.119945152	0	3	7
KDM5B	10	-2.440622041	1.672341898	-4.426264755	1.520627859	-2.644698125	1.472518803	-4.029747343	0.632268215	0.204076084	0.626915563	-1.233619677	0.888359643	1	6	3
KPNA4	10	-2.528758292	1.944900933	-5.807354922	-0.506959989	-3.013535648	1.104589434	-4.672425342	-1.915111102	0.484777356	1.890836071	-3.041820176	1.700439718	2	0	8
LRBA	10	-1.31579001	1.088423573	-2.925999419	-0.071790683	-1.85135077	1.3174493	-3.867896464	-0.289506617	0.53556076	0.367433346	-0.099368395	0.957816263	0	6	4
MAP2K4	10	1.169941256	1.791309781	-1.064130337	4.584962501	-0.25018695	1.179119514	-2.150559677	1.943416472	1.420128206	1.368631349	-0.358453971	3.814444347	0	3	7
MAPK6	10	-2.276639539	2.690140944	-5.087462841	3.624490865	-2.132898725	1.049107235	-3.169925001	0.447458977	-0.143740815	1.796780137	-2.685364398	3.177031888	5	1	4
MYCBP2	10	-1.856203802	3.190284503	-5.169925001	5.614709844	-1.436185426	2.044522917	-4.504344041	3.147643554	-0.420018376	1.631820641	-3	2.46706629	4	5	1
MYO5B	10	-2.154000386	1.60432532	-5.169925001	-0.180572246	-2.455860094	1.162597852	-4.426264755	-1.337441094	0.301859708	0.692392907	-0.743660247	1.309753381	1	5	4
PCDH1	10	-3.285812349	0.660161336	-4.217230716	-2.246311048	-4.567030003	1.265363703	-6.614709844	-2	1.281217655	1.417701843	-0.859678712	3.307281319	1	2	7
PLCG1	10	-2.497668702	1.428561233	-4.785724906	-0.362570079	-2.525231873	1.505400419	-5.502500341	-0.716207034	0.027563171	1.182785608	-2.321928095	1.349942471	3	2	5
RBBP6	10	-1.901308161	1.396394253	-3.736965594	1.115477217	-2.675715675	1.286577274	-5.108524457	-0.949373927	0.774407514	1.065446253	-0.399095955	2.845490051	0	5	5
SETDB1	10	-2.624938755	0.730119449	-3.736965594	-1.641105579	-2.427984223	1.004196944	-3.830074999	-0.637429921	-0.196954532	0.644182722	-1.432959407	0.508093737	2	8	0
SETX	10	-1.206123234	1.812268287	-4.478047297	2.362570079	-1.447674752	1.613561813	-3.471868722	2.192645078	0.241551519	0.615049226	-1.006178575	0.875723635	1	5	4
SRCAP	10	-2.444336214	0.893598916	-3.392317423	-0.506959989	-2.771552219	0.994991257	-4.569855608	-1.285402219	0.327216004	1.286690398	-1.047418122	3.447236322	2	6	2
SUV420H1	10	-2.786841439	1.423191712	-4.247927513	-0.700439718	-3.230860203	1.387303744	-4.68182404	-0.4639471	0.444018764	0.684718556	-0.236492618	2.053047431	0	8	2
TAOK1	10	-2.612194313	2.413108186	-5.273018494	1.222392421	-2.344811273	1.091026164	-3.475733431	-0.777607579	-0.26738304	1.93443182	-3.500428991	2	2	4	4
UBR4	10	-2.433560618	2.092898269	-6.442943496	-0.144389909	-2.326284655	1.021231857	-4.10433666	-0.844008804	-0.107275963	1.806585727	-4.161463423	1.662965013	2	4	4
USP39	10	-2.093306572	0.542096746	-2.906890596	-1.125530882	-4.899920381	1.452969477	-6.584962501	-2.36923381	2.806613809	1.578680852	0.107741569	5.266786541	0	1	9
UTRN	10	-2.456051374	1.261330114	-4.087462841	-0.980371193	-1.951497341	1.090517675	-3.312882955	-0.485426827	-0.504554032	1.802163661	-3.602036014	1.98550043	3	4	3
WDR43	10	-1.932099144	0.904033221	-2.807354922	0.085391491	-2.700107237	1.207811759	-4.672425342	-0.839535328	0.768008092	0.877086395	-0.426700417	2.534336428	0	5	5
ZNF587	10	-2.110480251	1.333479567	-4.397592365	-0.087462841	-4.456224006	2.038958839	-7.199672345	-1.719371534	2.345743755	1.456998527	0.297053426	4.943332592	0	2	8

Annex B

Table B.1: Distribution of *PIK3CA's* mutation impact on protein by MADE group

	METABRIC set			TCGA set		
	MADE_wt	mut=wt	MADE_mut	MADE_wt	mut=wt	MADE_mut
Oncogenic	6	39	39	21	75	45
Likely Neutral	0	3	2	1	4	4
No Anotated Alteration	0	5	0	1	7	3

Table B.2: P-values of pairwise test of distribution of *PIK3CA's* mutation impact on protein by MADE group (p-values are adjusted with Bonferroni's correction).

	METABRIC	TCGA
Oncogenic: Likely Neutral	1	1
Oncogenic: No Anotated Alteration	0.133	1
Likely Neutral: No Anotated Alteration	1	1

Annex C

Summary of METABRIC and TCGA set for *PIK3CA's* analysis

Factor	METABRIC set	TCGA set
Total Number	94	161
	Median (LQ, UQ)	*Median (LQ, UQ)*
Age at diagnosis (years)	60.61 (50.06, 69.62)	58.57 (49.82, 67.35)
Follow-up all cases (years)	8.99 (4.19, 13.75)	2.09(1.37, 5.13)
Follow-up still living (years)	12.16 (7.85, 17.36)	1.99 (1.30, 4.84)
Vital status		
Alive	47 (50%)	139 (86.3%)
Dead	47 (50%)	22 (13.7%)
Disease Specific Death	27 (28.7%)	11 (7%)
Tumour size	22 (17.18, 28.75)	—
NPI	4.04 (3.04, 5.05)	—
Lymph nodes positive		
Number (0, 1, 2, >3)	50, 11, 12, 21	—
Grade		
I	13 (13.8%)	—
II	35 (37.2%)	—
III	44 (46.8%)	—
ER status		
Positive	79 (84%)	145 (90%)
Negative	15 (16%)	16 (10%)
PR status		
Positive	67 (71.2%)	136 (84.5%)
Negative	27 (28.8%)	25 (15.5%)
HER2 status		
Positive	15 (16%)	93 (57.7%)
Negative	79 (84%)	22 (13.6%)
NA or Indeterminate	-	46 (28.7%)
Stage		
I	31 (33%)	29(18%)
II	27 (28.8%)	87 (54%)
III	8 (8.5%)	35 (21.8%)
IV	2 (2.1%)	5 (3.1%)
Not reported	26 (27.6%)	5 (3.1%)

Factor	METABRIC set	TCGA set
PAM50 subtype		
Basal	7 (7.4%)	7 (4.3%)
HER2	14 (14.9%)	10 (6.2%)
Luminal A	40 (42.6%)	110 (68.4%)
Luminal B	25 (26.6%)	30 (18.6%)
Normal	8 (8.5%)	4 (2.5%)
iCluster		
iC1	2 (2.1%)	—
iC2	6 (6.4%)	—
iC3	22 (23.4%)	—
iC4	13 (13.8%)	—
iC5	6 (6.4%)	—
iC6	2 (2.1%)	—
iC7	12 (12.8)	—
iC8	16 (17.0%)	—
iC9	10 (10.6%)	—
iC10	5 (5.3%)	—

Annex D

Table D.1: Summary of *PIK3CA's* α, β and γ ratios for METABRIC and TCGA sets

	METABRIC					
	min	**1st Q**	**median**	**mean**	**3rd Q**	**max**
α	-2.4681	-1.5495	-0.9558	-0.7468	-0.1778	5.7549
β	-3.6088	-1.9481	-1.3703	-1.2949	-0.9471	5.1657
γ	-1.21501	0.02915	0.48048	0.54809	0.95779	3.70191

	TCGA					
	min	**1st Q**	**median**	**mean**	**3rd Q**	**max**
α	-4.8244	-1.6189	-0.7975	-0.8019	-0.0555	7.7814
β	-5.5661	-1.8074	-1.0356	-1.0665	-0.4274	5.8209
γ	-2.102	-0.2314	0.1982	0.2646	0.7417	2.4557

Table D.2: P-values of pairwise comparison between *PIK3CA's* α, β and γ ratios means for METABRIC and TCGA sets (p-values are adjusted with Bonferroni's correction)

	METABRIC	**TCGA**
α: β	0.00033	0.19
α: γ	<2e-16	<2e-16
β: γ	<2e-16	<2e-16

Annex E – Summary of METABRIC *PIK3CA's* Survival Analysis

Table E.1: Summary Statistic of *PIK3CA's* MADE Survival Analysis in METABRIC set (Median, LCL and UCL in days; LCL: lower confidence level; UCL: uper confidence level; N: number of people of each group).

		Overall Survival					Disease Specific Survival				
		N	**Events**	**Median**	**0.95LCL**	**0.95UCL**	**N**	**Events**	**Median**	**0.95LCL**	**0.95UCL**
All	MADE	45	26	2907	2112	NA	45	18	NA	2582	NA
	mut=wt	46	21	6750	4139	NA	46	9	NA	NA	NA
ER positive	MADE	33	15	4101	3447	NA	33	9	NA	NA	NA
	mut=wt	42	18	6750	4343	NA	42	8	NA	NA	NA
ER negative	MADE	12	11	1028	939	NA	12	9	1374	962	NA
	mut=wt	4	3	2539	744	NA	4	1	NA	744	NA
PR positive	MADE	27	13	4101	3447	NA	27	8	NA	3617	NA
	mut=wt	37	14	6750	4680	NA	37	7	NA	NA	NA
PR negative	MADE	18	13	1456	962	NA	18	10	2149	985	NA
	mut=wt	9	7	3660	1296	NA	9	2	NA	1799	NA
HER2 positive	MADE	9	6	2149	1071	NA	9	5	2149	1530	NA
	mut=wt	6	3	4138	2048	NA	6	2	NA	2048	NA
HER2 negative	MADE	36	20	3617	2697	NA	36	13	NA	3447	NA
	mut=wt	40	18	6750	4297	NA	40	7	NA	NA	NA

Table E.2: Summary Statistic of *PIK3CA's* MADE groups Survival Analysis in METABRIC set (Median, LCL and UCL in days; LCL: lower confidence level; UCL: uper confidence level; N: number of people of each group).

		Overall Survival					Disease Specific Survival				
		N	**Events**	**Median**	**0.95LCL**	**0.95UCL**	**N**	**Events**	**Median**	**0.95LCL**	**0.95UCL**
All	MADE_wt	6	4	1382	1293	NA	6	3	1382	1293	NA
	mut=wt	46	21	6750	4138	NA	46	9	NA	NA	NA
	MADE_mut	39	22	3617	2149	NA	39	15	NA	2907	NA
ER positive	MADE_wt	5	3	2737	1293	NA	5	2	NA	1293	NA
	mut=wt	42	18	6750	4341	NA	42	8	NA	NA	NA
	MADE_mut	28	12	4550	3617	NA	18	7	NA	NA	NA
ER negative	MADE_wt	1	1	1382	NA	NA	1	1	1382	NA	NA
	mut=wt	4	3	2539	744	NA	4	1	NA	744	NA
	MADE_mut	11	10	985	939	NA	11	8	1374	939	NA
PR positive	MADE_wt	5	3	2737	1293	NA	5	2	NA	1293	NA
	mut=wt	37	14	6750	4680	NA	37	7	NA	NA	NA
	MADE_mut	22	10	4550	3617	NA	22	6	NA	3617	NA
PR negative	MADE_wt	1	1	1382	NA	NA	1	1	1382	NA	NA
	mut=wt	9	7	3660	1296	NA	9	2	NA	1799	NA
	MADE_mut	17	12	1530	962	NA	17	9	2149	985	NA
HER2 positive	MADE_wt	0					0	0			
	mut=wt	6	3	4138	2048	NA	6	2	NA	2048	NA
	MADE_mut	9	6	2149	1071	NA	9	5	2149	1530	NA
HER2 negative	MADE_wt	6	4	1382	1293	NA	6	3	1382	1293	NA
	mut=wt	40	18	6750	4277	NA	40	7	NA	NA	NA
	MADE_mut	30	16	3950	2907	NA	30	10	NA	3447	NA

Table E.3: Summary Statistic of *PIK3CA's* α_DAE groups Survival Analysis in METABRIC set (Median, LCL and UCL in days; LCL: lower confidence level; UCL: uper confidence level; N: number of people of each group).

		Overall Survival					Disease Specific Survival				
		N	**Events**	**Median**	**0.95LCL**	**0.95UCL**	**N**	**Events**	**Median**	**0.95LCL**	**0.95UCL**
All	α_DAE	69	35	4138	3447	NA	69	19	NA	NA	NA
	α_noDAE	22	12	4550	1799	NA	22	8	NA	2582	NA
ER positive	α_DAE	57	25	6750	3681	NA	57	12	NA	NA	NA
	α_noDAE	18	8	5280	3950	NA	18	5	NA	5280	NA
ER negative	α_DAE	12	10	974	939	NA	12	7	1374	962	NA
	α_noDAE	4	4	1226	125	NA	4	3	1382	125	NA
PR positive	α_DAE	50	21	6750	4101	NA	50	12	NA	NA	NA
	α_noDAE	14	6	5280	3950	NA	14	3	NA	5280	NA
PR negative	α_DAE	19	14	2149	962	NA	19	7	NA	9855	NA
	α_noDAE	8	6	1664	1382	NA	8	5	1799	1382	NA
HER2 positive	α_DAE	10	5	4138	2048	NA	10	4	NA	2048	NA
	α_noDAE	5	4	1530	1071	NA	5	3	2582	1530	NA
HER2 negative	α_DAE	59	30	4277	3447	NA	59	15	NA	NA	NA
	α_noDAE	17	8	5280	2112	NA	17	5	NA	5280	NA

Table E.4: Summary Statistic of *PIK3CA's* α_DAE groups Survival Analysis in METABRIC set (Median, LCL and UCL in days; LCL: lower confidence level; UCL: uper confidence level; N: number of people of each group).

		Overall Survival					Disease Specific Survival				
		N	**Events**	**Median**	**0.95LCL**	**0.95UCL**	**N**	**Events**	**Median**	**0.95LCL**	**0.95UCL**
All	α_DAEwt	61	30	4277	3617	NA	61	15	NA	NA	NA
	α_noDAE	22	12	4550	1799	NA	22	8	NA	2582	NA
	α_DAEmut	8	5	1722	985	NA	8	4	2149	985	NA
ER positive	α_DAEwt	54	24	4680	3681	NA	54	12	NA	NA	NA
	α_noDAE	18	8	5280	3950	NA	18	5	NA	5280	NA
	α_DAEmut	3	1	NA	1296	NA	3	0	NA	NA	NA
ER negative	α_DAEwt	7	6	940	744	NA	7	3	1374	939	NA
	α_noDAE	4	4	1226	125	NA	4	3	1382	125	NA
	α_DAEmut	5	4	985	962	NA	5	4	985	962	NA
PR positive	α_DAEwt	49	21	4680	3681	NA	49	12	NA	NA	NA
	α_noDAE	14	6	5280	3950	NA	14	3	NA	5280	NA
	α_DAEmut	1	0	NA	NA	NA	1	0	NA	NA	NA
PR negative	α_DAEwt	12	9	2697	939	NA	12	3	NA	NA	NA
	α_noDAE	8	6	1664	1382	NA	8	5	1799	1382	NA
	α_DAEmut	7	5	1296	962	NA	7	4	2149	962	NA
HER2 positive	α_DAEwt	6	3	4138	2048	NA	6	2	NA	2048	NA
	α_noDAE	5	4	1530	1071	NA	5	3	2582	1530	NA
	α_DAEmut	4	2	2149	962	NA	4	2	2149	962	NA
HER2 negative	α_DAEwt	55	27	4341	3617	NA	57	13	NA	NA	NA
	α_noDAE	17	8	5280	2112	NA	17	5	NA	5280	NA
	α_DAEmut	4	3	1140	839	NA	4	2	985	839	NA

Annex F – *PIK3CA's* Kaplan-Meier Survival Analysis of METABRIC set

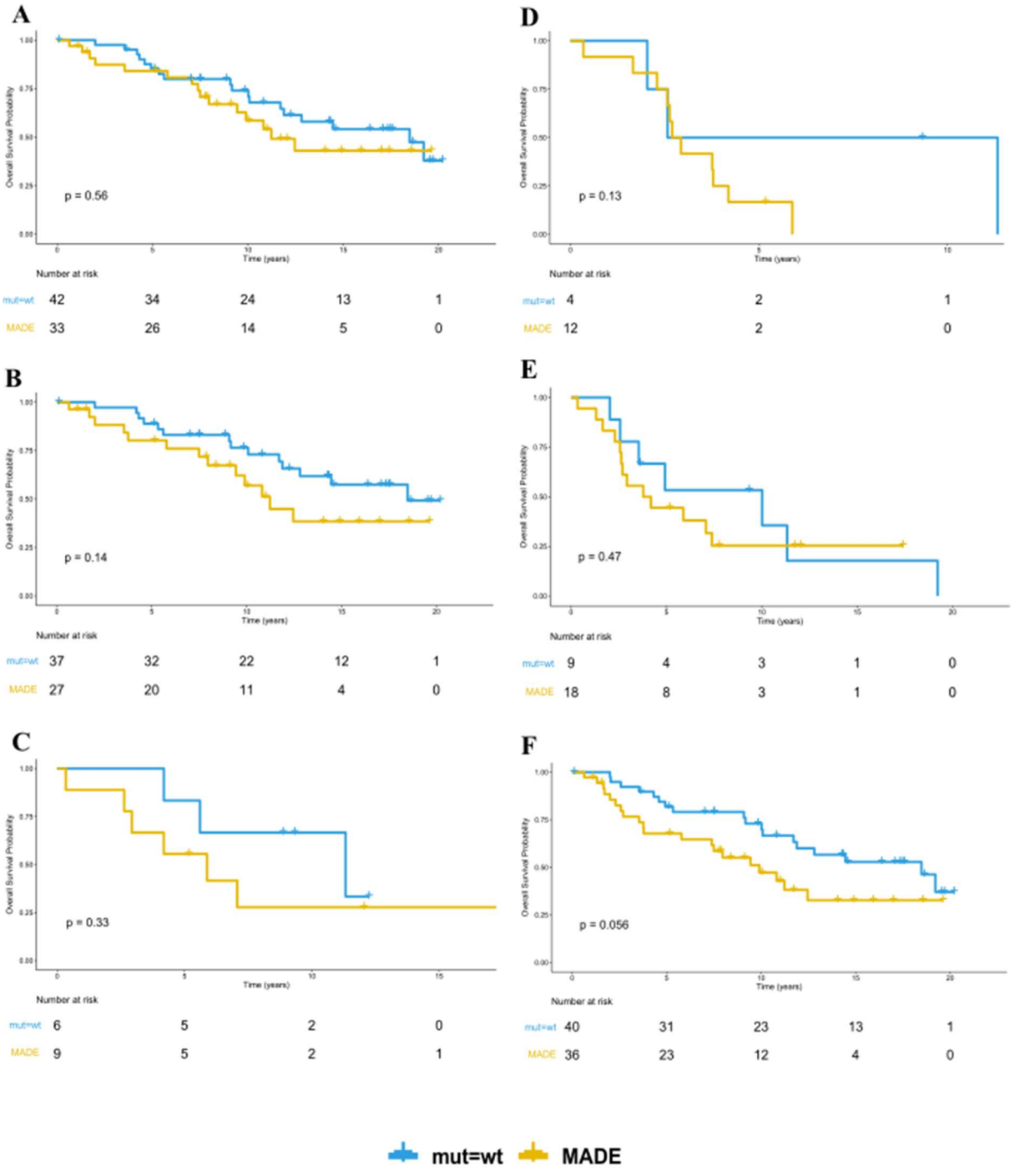


Figure F.1: Overall survival of *PIK3CA's* MADE in METABRIC set according to ER, PR and HER2 statuses. Overall survival in (A) ER positive group, (B) PR positive group, (C) HER2 positive group, (D) ER negative group, (E) PR negative group and (F) HER2 negative group.

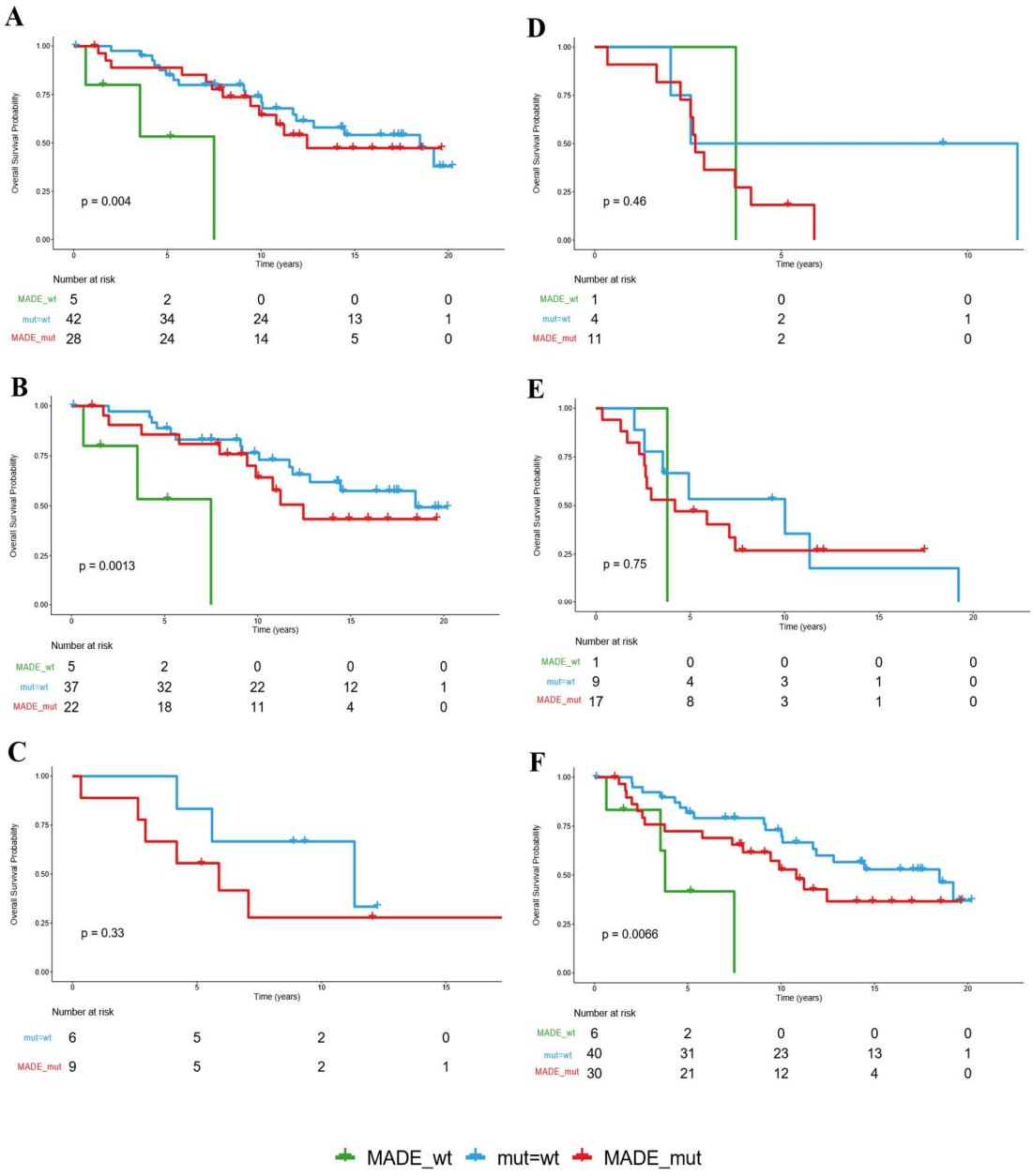


Figure F.2: Overall survival of *PIK3CA's* MADE groups in METABRIC set according to ER, PR and HER2 statuses. Overall survival in (A) ER positive group, (B) PR positive group, (C) HER2 positive group, (D) ER negative group, (E) PR negative group and (F) HER2 negative group.

Table F.1: P-values of pairwise comparison of overall survival of *PIK3CA's* MADE in METABRIC set (p-values calculated by log-rank test if curves do not cross and by Two-Stage procedure if curves cross).

	ER positive	ER negative	PR positive	PR negative	HER2 positive	HER2 negative
MADE_wt : mut=wt	0.0012	0.18	0.00053	0.7437		0.00074
MADE_wt : MADE_mut	0.0074	0.21	0.0071	0.0313	0.33	0.7298
mut=wt : MADE_mut	0.8569	0.2524	0.93352	0.384		0.3463

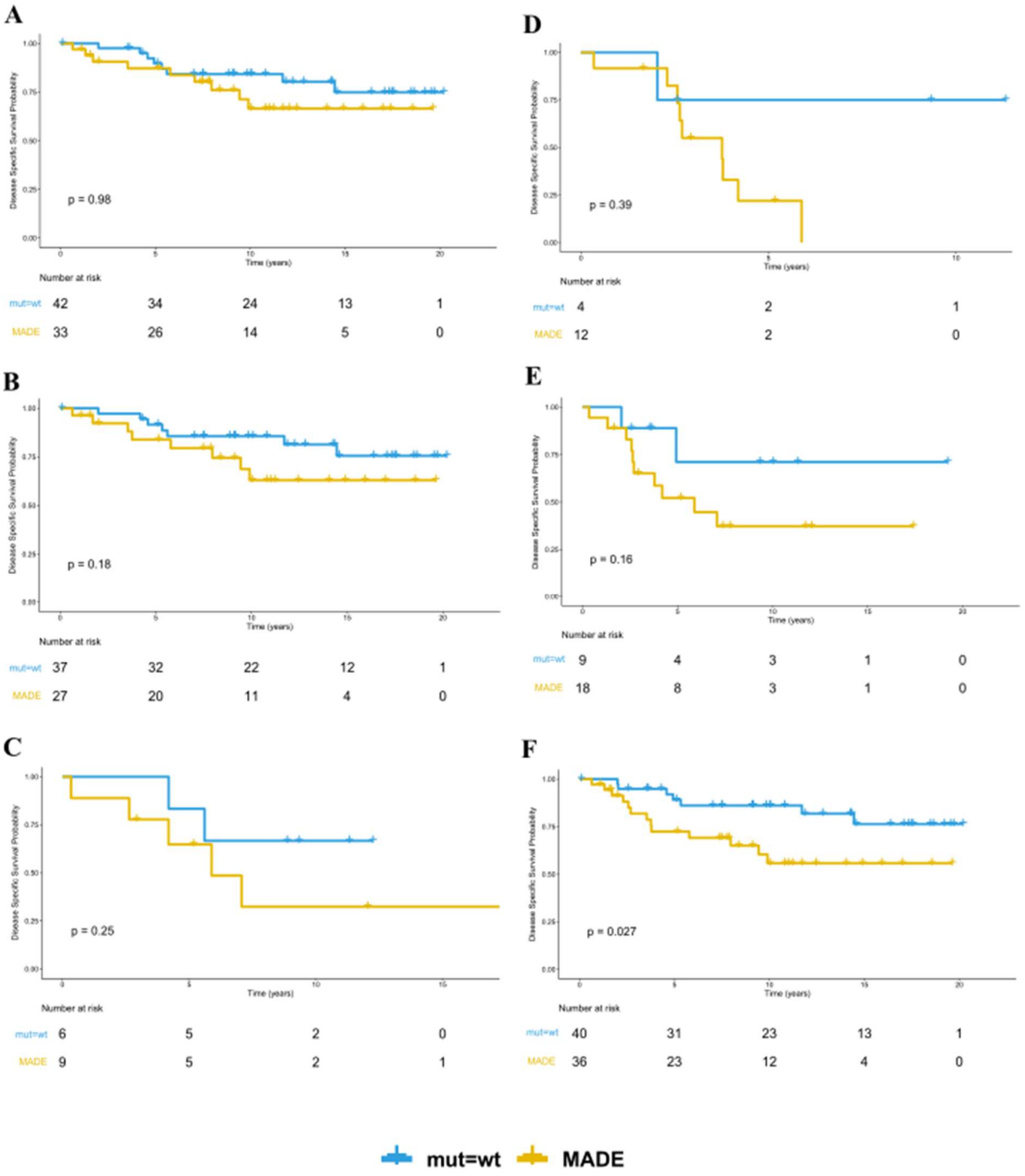


Figure F.3: Disease specific survival of *PIK3CA's* MADE in METABRIC set according to ER, PR and HER2 statuses. Disease specific survival in (A) ER positive group, (B) PR positive group, (C) HER2 positive group, (D) ER negative group, (E) PR negative group and (F) HER2 negative group.

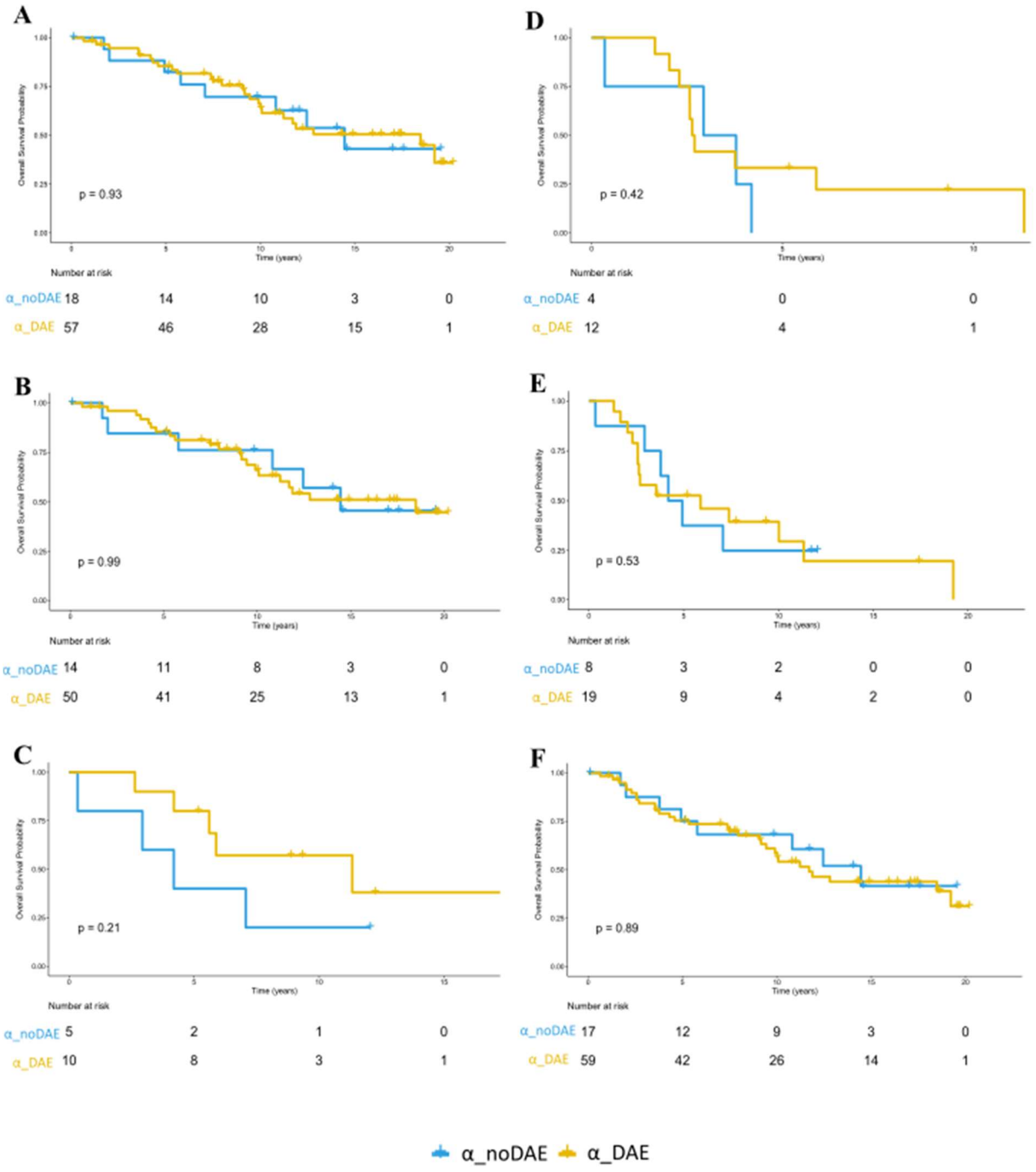


Figure F.4: Overall survival of *PIK3CA's* α_DAE groups in METABRIC set according to ER, PR and HER2 statuses. Overall survival in (A) ER positive group, (B) PR positive group, (C) HER2 positive group, (D) ER negative group, (E) PR negative group and (F) HER2 negative group.

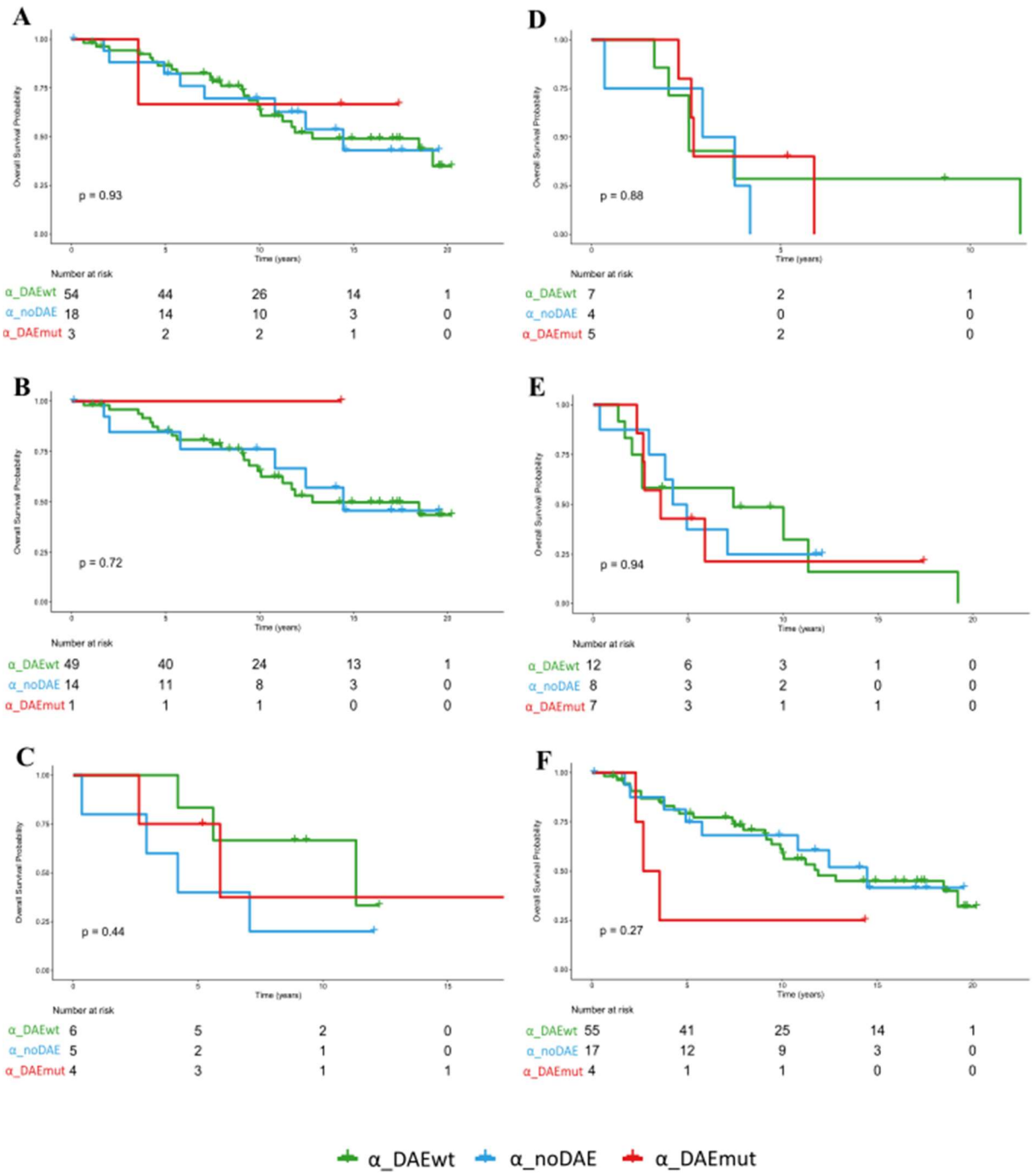


Figure F.5: Overall survival of α_DAE groups in METABRIC set according to ER, PR and HER2 statuses. Overall survival in (A) ER positive group, (B) PR positive group, (C) HER2 positive group, (D) ER negative group, (E) PR negative group and (F) HER2 negative group.

Table F.2: P-values of pairwise comparison of overall survival of *PIK3CA's* α_DAE groups in METABRIC set (p-values calculated by log-rank test if curves do not cross and by Two-Stage procedure if curves cross).

	ER positive	ER negative	PR positive	PR negative	HER2 positive	HER2 negative
α_DAEwt : α_noDAE	0.8632	0.715	0.8503	0.4125	0.1722	0.9683
α_DAEwt : α_DAEmut	0.0371	0.374	0.0255	0.1383	0.697	0.1367
α_noDAE : α_DAEmut	0.7543	0.9633	0.0732	0.4909	0.5337	0.6313

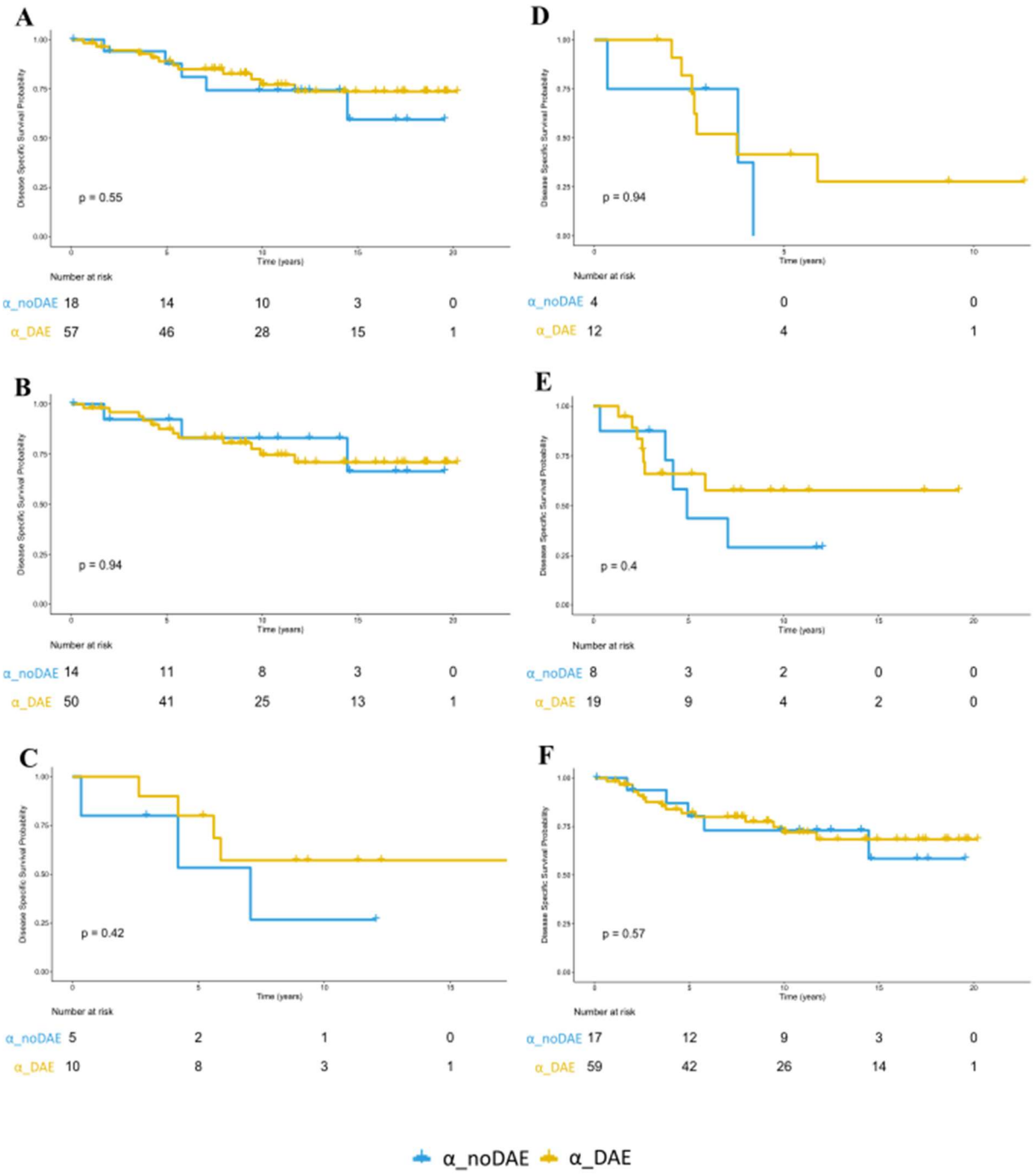


Figure F.6: Disease specific survival of *PIK3CA's* α_DAE group in METABRIC set according to ER, PR and HER2 statuses. Disease specific survival in (A) ER positive group, (B) PR positive group, (C) HER2 positive group, (D) ER negative group, (E) PR negative group and (F) HER2 negative group.

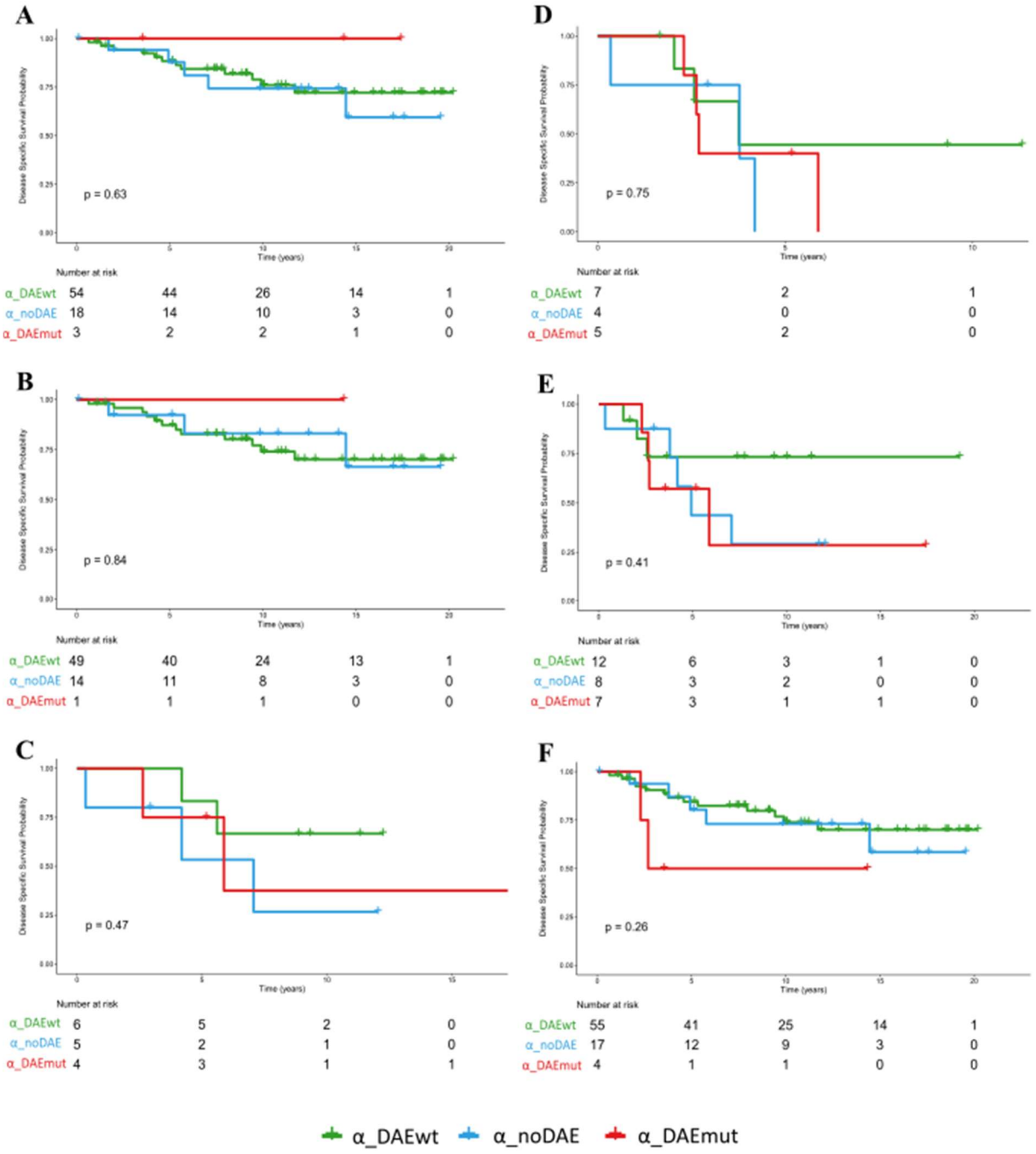


Figure F.7: Disease specific survival of *PIK3CA's* α_DAE groups in METABRIC set according to ER, PR and HER2 statuses. Disease specific survival in (A) ER positive group, (B) PR positive group, (C) HER2 positive group, (D) ER negative group, (E) PR negative group and (F) HER2 negative group.

Table F.3: P-values of pairwise comparison of disease specific survival of *PIK3CA's* α_DAE groups in METABRIC set (p-values calculated by log-rank test if curves do not cross and by Two-Stage procedure if curves cross).

	ER positive	ER negative	PR positive	PR negative	HER2 positive	HER2 negative
α_DAEwt : α_noDAE	0.5709	0.6031	0.9265	0.8617	0.1773	0.5873
α_DAEwt : α_DAEmut	0.4029	0.7521	0.5547	0.6164	0.4756	0.824
α_noDAE : α_DAEmut	0.36	0.9186	0.674	0.646	0.751	0.4158

Annex G

Table G.1: Summary Statistic of *PIK3CA's* MADE Survival Analysis in TCGA set (Median, LCL and UCL in days; LCL: lower confidence level; UCL: uper confidence level; N: number of people of each group).

		Overall Survival					Disease Specific Survival				
		N	**Events**	**Median**	**0.95LCL**	**0.95UCL**	**N**	**Events**	**Median**	**0.95LCL**	**0.95UCL**
All	MADE	86	10	3409	3262	NA	73	6	NA	NA	NA
	mut=wt	75	12	4267	4267	NA	85	5	NA	3360	NA
ER positive	MADE	64	10	4267	4267	NA	62	4	NA	NA	NA
	mut=wt	81	9	3409	3262	NA	80	4	NA	3360	NA
ER negative	MADE	11	2	NA	1993	NA	11	2	NA	1965	NA
	mut=wt	5	1	NA	571	NA	5	1	NA	564	NA
PR positive	MADE	59	9	4267	4267	NA	57	4	NA	NA	NA
	mut=wt	77	7	3409	3262	NA	76	3	NA	3360	NA
PR negative	MADE	19	3	NA	1993	NA	16	2	NA	1965	NA
	mut=wt	9	3	792	571	NA	9	2	NA	780	NA
HER2 positive	MADE	11	0	NA	NA	NA	11	0	NA	NA	NA
	mut=wt	11	3	3409	NA	NA	11	2	3360	NA	NA
HER2 negative	MADE	44	5	NA	NA	NA	43	3	NA	NA	NA
	mut=wt	49	2	NA	NA	NA	49	0	NA	NA	NA

Table G.2: Summary Statistic of *PIK3CA's* MADE groups Survival Analysis in TCGA set (Median, LCL and UCL in days; LCL: lower confidence level; UCL: uper confidence level; N: number of people of each group).

		Overall Survival					Disease Specific Survival				
		N	**Events**	**Median**	**0.95LCL**	**0.95UCL**	**N**	**Events**	**Median**	**0.95LCL**	**0.95UCL**
All	MADE_wt	23	3	NA	NA	NA	23	2	NA	NA	NA
	mut=wt	86	0	3409	3262	NA	85	5	NA	3360	NA
	MADE_mut	52	9	4267	4267	NA	50	4	NA	NA	NA
ER positive	MADE_wt	22	3	NA	1884	NA	22	2	NA	NA	NA
	mut=wt	81	9	3409	3262	NA	80	4	NA	3360	NA
	MADE_mut	42	7	4267	4267	NA	40	2	NA	NA	NA
ER negative	MADE_wt	1	0	NA	NA	NA	1	0	NA	NA	NA
	mut=wt	5	1	NA	571	NA	5	1	NA	564	NA
	MADE_mut	10	2	1993	548	NA	10	2	1965	564	NA
PR positive	MADE_wt	22	3	NA	1884	NA	22	2	NA	NA	NA
	mut=wt	77	7	3409	3262	NA	76	3	NA	3360	NA
	MADE_mut	37	6	4267	2348	NA	35	2	NA	NA	NA
PR negative	MADE_wt	1	0	NA	NA	NA	1	0	NA	NA	NA
	mut=wt	9	3	792	571	NA	9	2	NA	780	NA
	MADE_mut	15	3	NA	1993	NA	15	2	NA	1965	NA
HER2 positive	MADE_wt	1	0	NA	NA	NA	1	0	NA	NA	NA
	mut=wt	11	3	3409	NA	NA	11	2	3360	NA	NA
	MADE_mut	10	0	NA	NA	NA	10	0	NA	NA	NA
HER2 negative	MADE_wt	17	3	NA	1884	NA	17	2	NA	1857	NA
	mut=wt	49	2	NA	NA	NA	49	0	NA	NA	NA
	MADE_mut	27	2	NA	NA	NA	26	1	NA	NA	NA

Table G.3: Summary Statistic of *PIK3CA's* α_DAE groups Survival Analysis in TCGA set (Median, LCL and UCL in days; LCL: lower confidence level; UCL: uper confidence level; N: number of people of each group).

		Overall Survival					Disease Specific Survival				
		N	**Events**	**Median**	**0.95LCL**	**0.95UCL**	**N**	**Events**	**Median**	**0.95LCL**	**0.95UCL**
All	α_DAE	108	13	4267	3262	NA	106	4	NA	NA	NA
	α_noDAE	53	9	3409	1993	NA	52	7	3360	3360	NA
ER positive	α_DAE	98	12	4267	3262	NA	96	3	NA	NA	NA
	α_noDAE	47	7	3409	3409	NA	46	5	NA	3360	NA
ER negative	α_DAE	10	1	NA	NA	NA	10	1	NA	NA	NA
	α_noDAE	6	2	1993	571	NA	6	2	1965	564	NA
PR positive	α_DAE	91	10	4267	3262	NA	89	2	NA	NA	NA
	α_noDAE	45	6	NA	3409	NA	44	5	NA	3360	NA
PR negative	α_DAE	7	3	NA	792	NA	17	2	NA	NA	NA
	α_noDAE	8	3	1993	571	NA	8	2	1965	1965	NA
HER2 positive	α_DAE	19	1	NA	NA	NA	13	0	NA	NA	NA
	α_noDAE	9	2	3409	NA	NA	9	2	3360	NA	NA
HER2 negative	α_DAE	64	4	NA	NA	NA	64	1	NA	NA	NA
	α_noDAE	29	3	NA	NA	NA	28	2	NA	NA	NA

Table G.4: Summary Statistic of *PIK3CA's* α_DAE groups Survival Analysis in TCGA set (Median, LCL and UCL in days; LCL: lower confidence level; UCL: uper confidence level; N: number of people of each group).

		Overall Survival					Disease Specific Survival				
		N	**Events**	**Median**	**0.95LCL**	**0.95UCL**	**N**	**Events**	**Median**	**0.95LCL**	**0.95UCL**
All	α_DAEwt	90	10	4267	3262	NA	89	3	NA	NA	NA
	α_noDAE	53	9	3409	1993	NA	52	7	3360	3360	NA
	α_DAEmut	18	3	NA	NA	NA	17	1	NA	NA	NA
ER positive	α_DAEwt	85	10	4267	3262	NA	84	3	NA	NA	NA
	α_noDAE	47	7	3409	3409	NA	46	5	NA	3360	NA
	α_DAEmut	13	2	NA	NA	NA	12	0	NA	NA	NA
ER negative	α_DAEwt	5	0	NA	NA	NA	5	0	NA	NA	NA
	α_noDAE	6	2	1993	571	NA	6	2	1965	564	NA
	α_DAEmut	5	1	NA	548	NA	5	1	NA	564	NA
PR positive	α_DAEwt	81	9	4267	3262	NA	80	2	NA	NA	NA
	α_noDAE	45	6	NA	3409	NA	44	5	NA	3360	NA
	α_DAEmut	10	1	NA	NA	NA	9	0	NA	NA	NA
PR negative	α_DAEwt	9	1	NA	792	NA	9	1	NA	780	NA
	α_noDAE	8	3	1993	571	NA	8	2	1965	1965	NA
	α_DAEmut	8	2	NA	723	NA	8	1	NA	NA	NA
HER2 positive	α_DAEwt	10	1	NA	NA	NA	10	0	NA	NA	NA
	α_noDAE	9	2	3409	NA	NA	9	2	3360	NA	NA
	α_DAEmut	3	0	NA	NA	NA	3	0	NA	NA	NA
HER2 negative	α_DAEwt	54	4	NA	NA	NA	54	1	NA	NA	NA
	α_noDAE	29	3	NA	NA	NA	28	2	NA	NA	NA
	α_DAEmut	10	0	NA	NA	NA	10	0	NA	NA	NA

Annex H – *PIK3CA's* Kaplan-Meier Survival Analysis of TCGA set

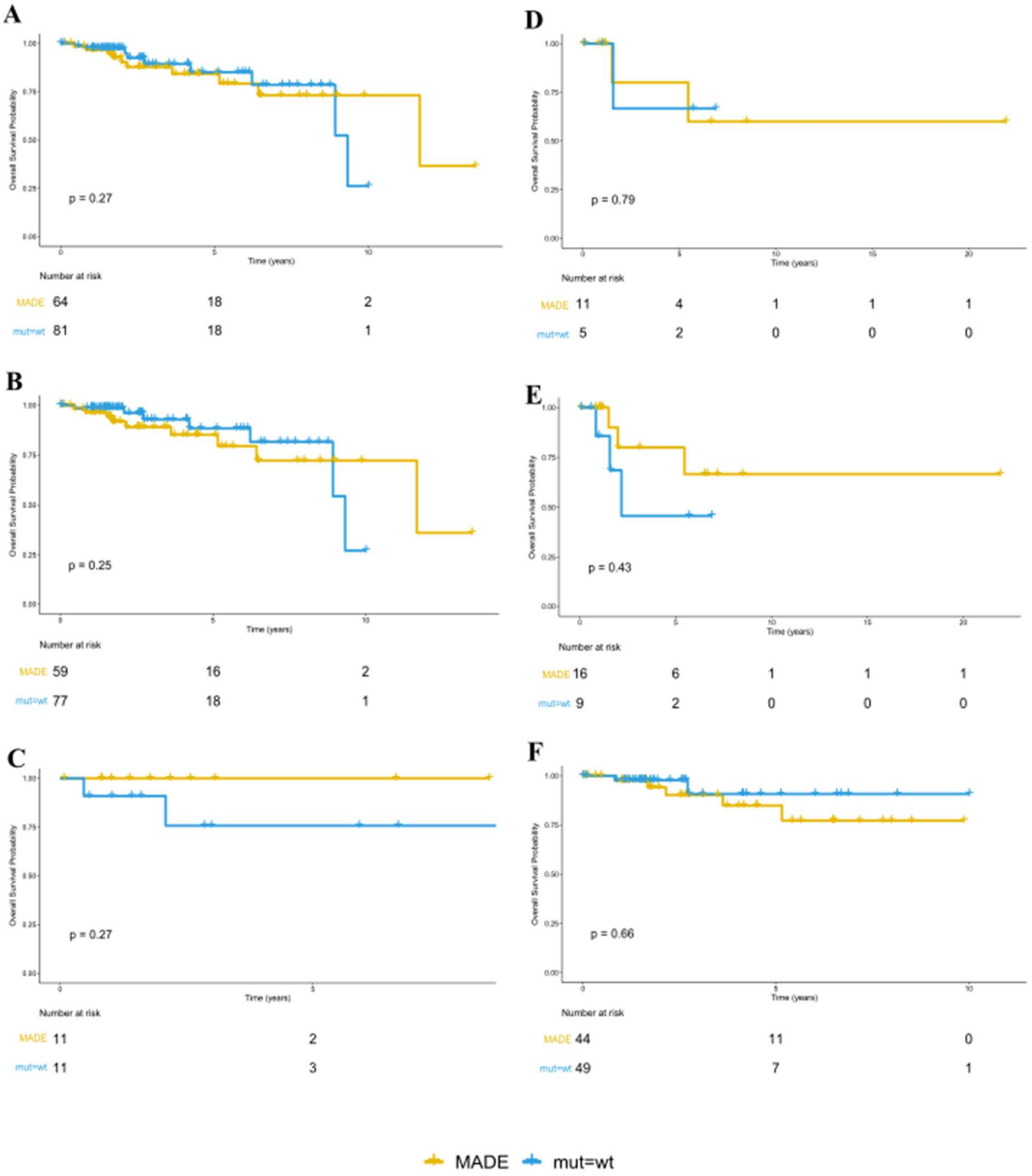


Figure H.1: Overall survival of *PIK3CA's* MADE in TCGA set according to ER, PR and HER2 statuses. Overall survival in (A) ER positive group, (B) PR positive group, (C) HER2 positive group, (D) ER negative group, (E) PR negative group and (F) HER2 negative group.

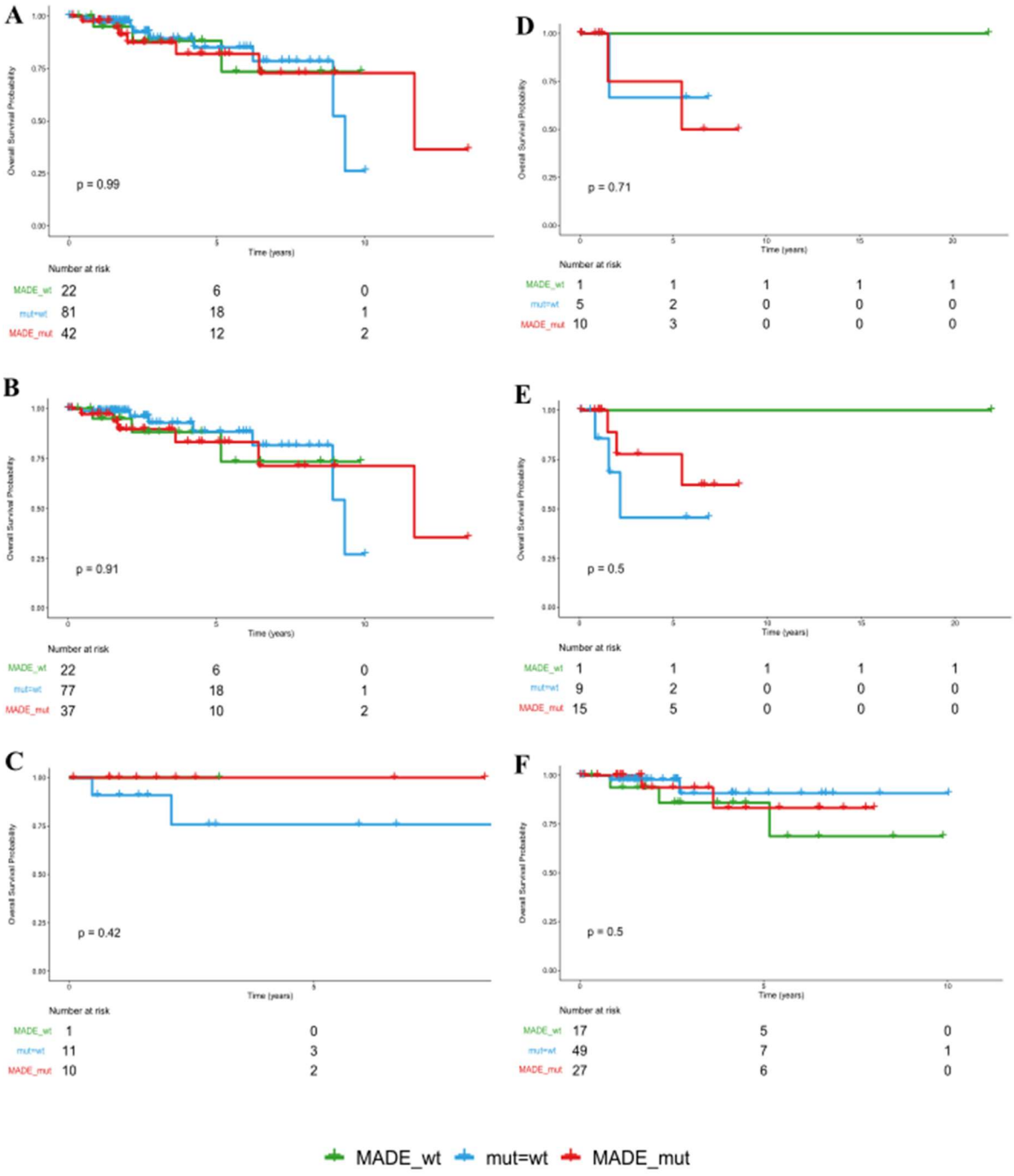


Figure H.2: Overall survival of *PIK3CA's* MADE groups in TCGA set according to ER, PR and HER2 statuses. Overall survival in (A) ER positive group, (B) PR positive group, (C) HER2 positive group, (D) ER negative group, (E) PR negative group and (F) HER2 negative group.

Table H.1: P-values of pairwise comparison of overall survival of *PIK3CA's* MADE in TCGA set (p-values calculated by log-rank test if curves do not cross and by Two-Stage procedure if curves cross).

	ER positive	ER negative	PR positive	PR negative	HER2 positive	HER2 negative
MADE_wt : mut=wt	0.5784	0.6653	0.6249		0.8621	0.2927
MADE_wt : MADE_mut	0.8981	0.4452	0.9275	0.6	—	0.5557
mut=wt : MADE_mut	0.3121	0.7015	0.2795		0.2687	0.7324

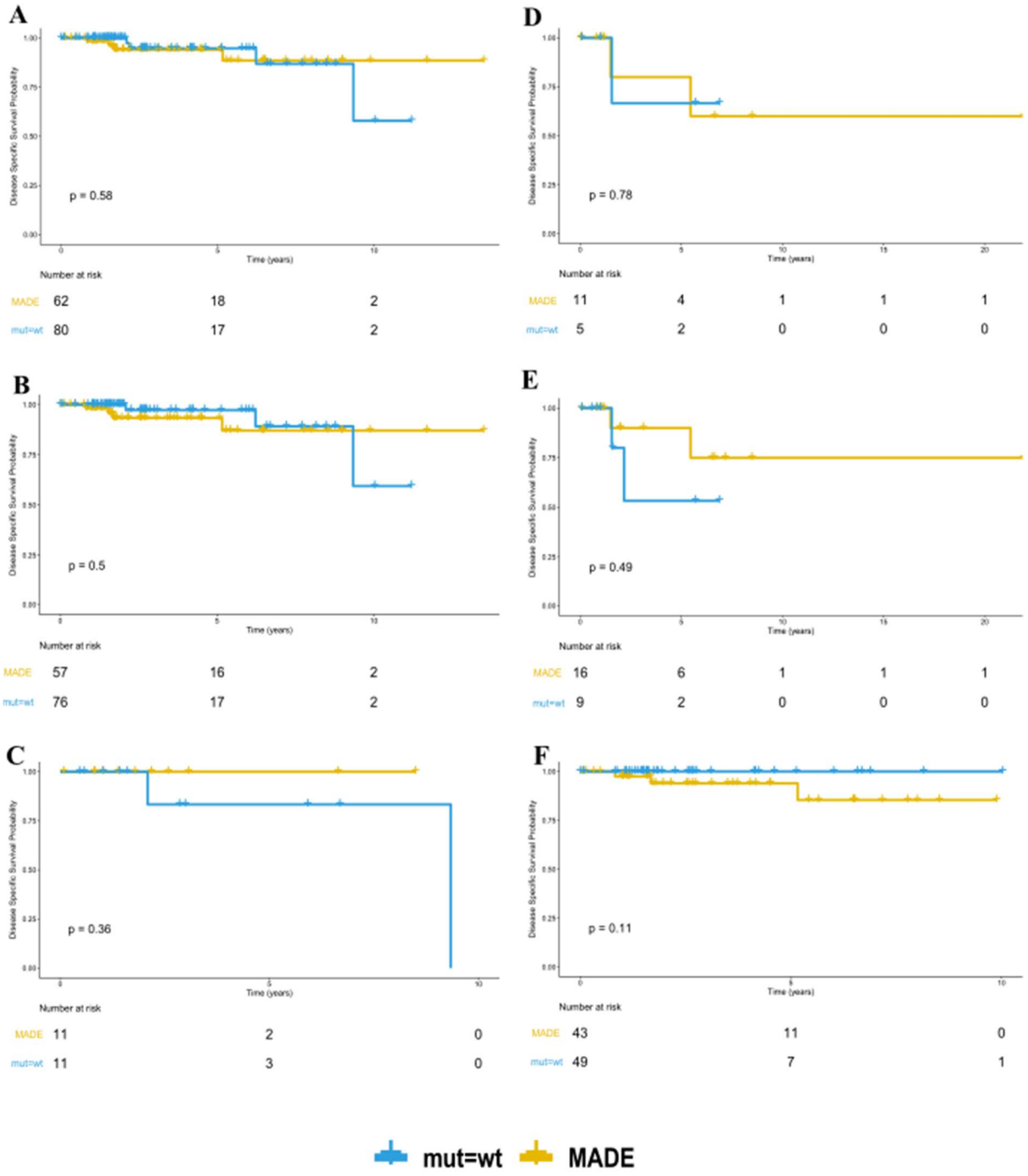


Figure H.3: Disease specific survival of *PIK3CA's* MADE in TCGA set according to ER, PR and HER2 statuses. Disease specific survival in (A) ER positive group, (B) PR positive group, (C) HER2 positive group, (D) ER negative group, (E) PR negative group and (F) HER2 negative group.

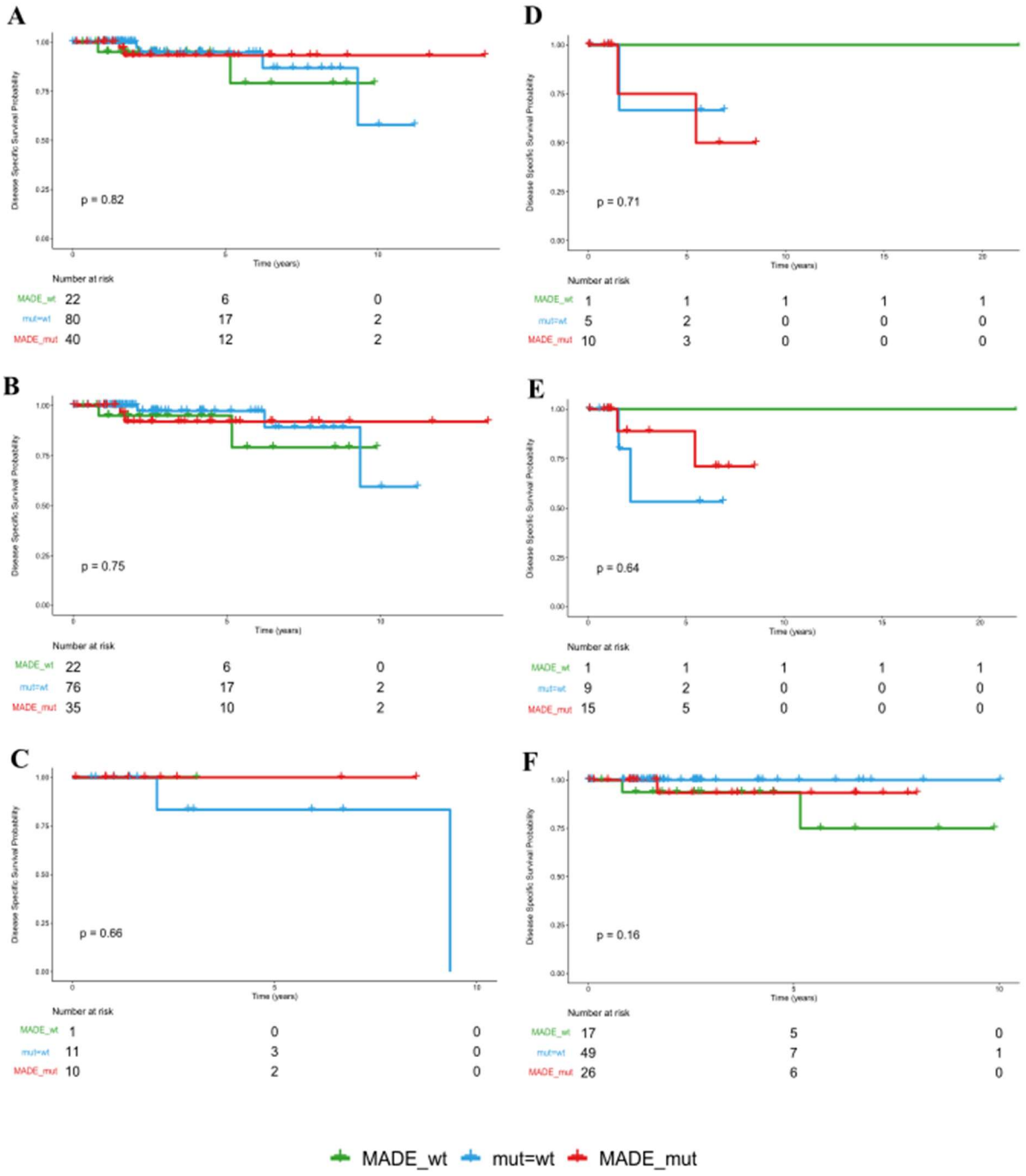


Figure H.4: Disease specific survival of *PIK3CA's* MADE groups in TCGA set according to ER, PR and HER2 statuses. Disease specific survival in (A) ER positive group, (B) PR positive group, (C) HER2 positive group, (D) ER negative group, (E) PR negative group and (F) HER2 negative group.

Table H.2: P-values of pairwise comparison of disease specific survival of *PIK3CA's* MADE in TCGA set (p-values calculated by log-rank test if curves do not cross and by Two-Stage procedure if curves cross).

	ER positive	ER negative	PR positive	PR negative	HER2 positive	HER2 negative
MADE_wt : mut=wt	0.6838	0.6625	0.5218	0.2261		0.0534
MADE_wt : MADE_mut	0.632	0.4452	0.7456	0.5772	0.96	0.5418
mut=wt : MADE_mut	0.9234	0.6311	0.8006	0.5605		0.2059

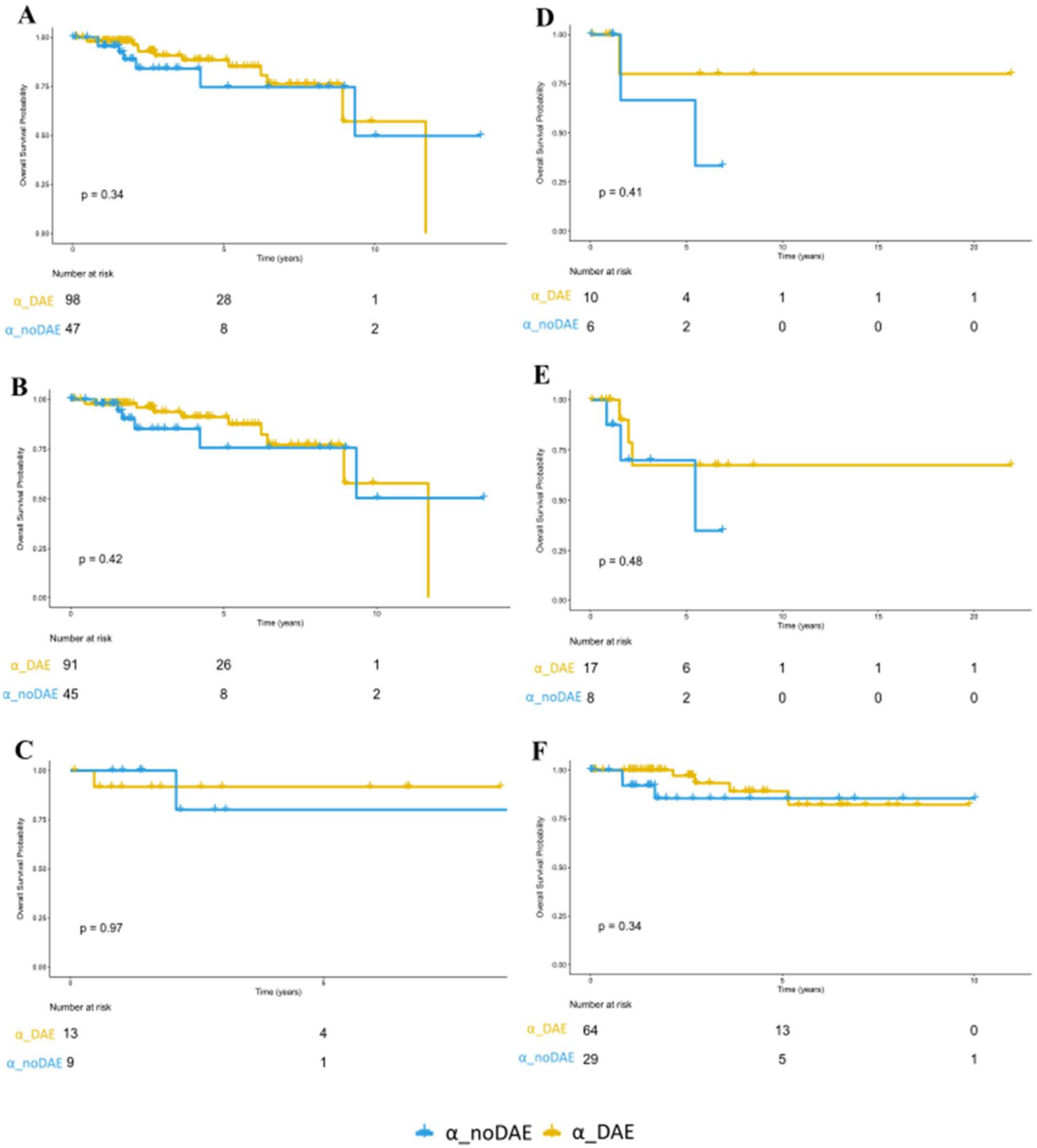


Figure H.5: Overall survival of *PIK3CA's* α_DAE groups in TCGA set according to ER, PR and HER2 statuses. Overall survival in (A) ER positive group, (B) PR positive group, (C) HER2 positive group, (D) ER negative group, (E) PR negative group and (F) HER2 negative group.

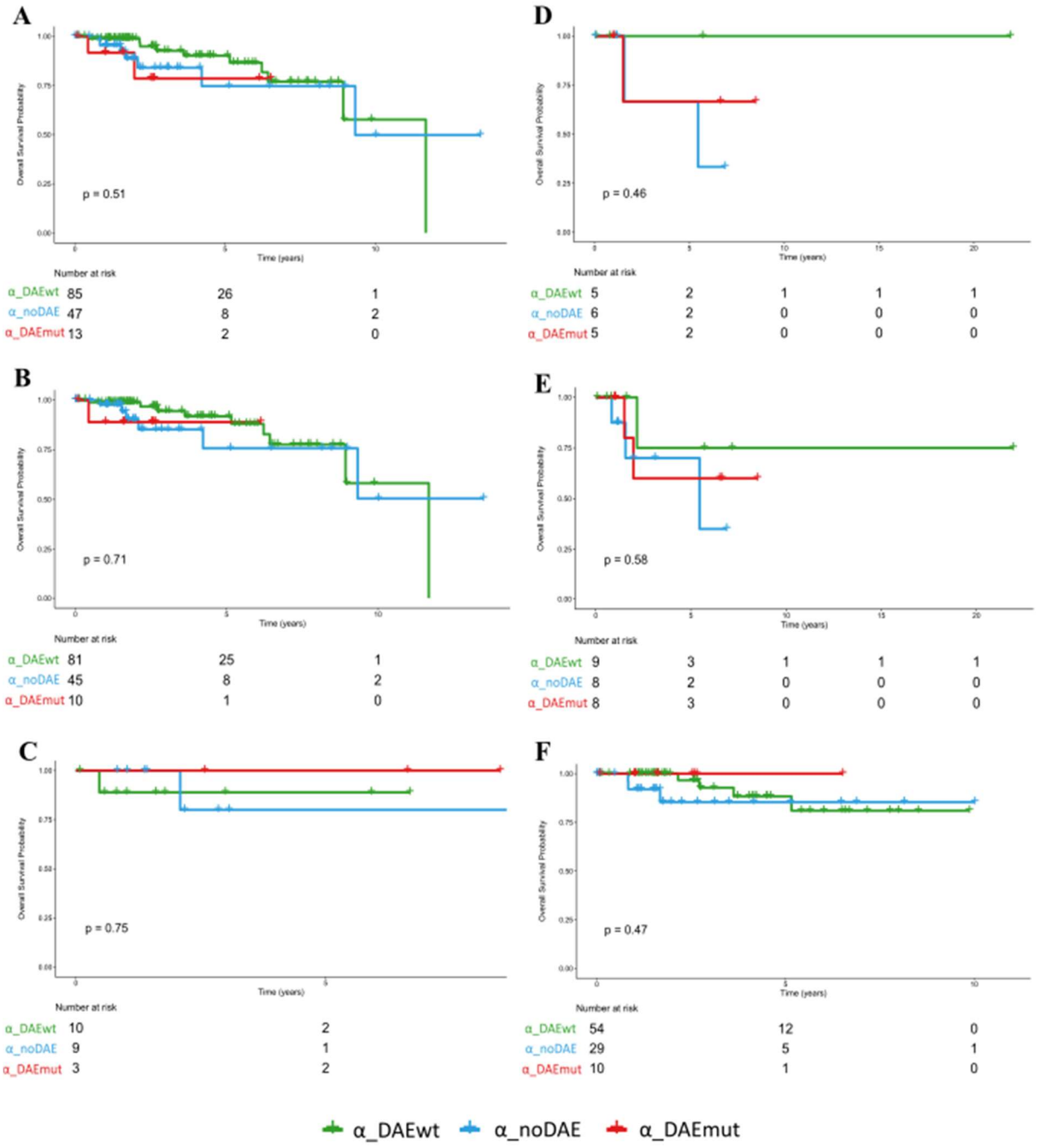


Figure H.6: Overall survival of *PIK3CA's* α_DAE groups in TCGA set according to ER, PR and HER2 statuses. Overall survival in (A) ER positive group, (B) PR positive group, (C) HER2 positive group, (D) ER negative group, (E) PR negative group and (F) HER2 negative group.

Table H.3: P-values of pairwise comparison of overall survival of PIK3CA's α_DAE groups in TCGA set (p-values calculated by log-rank test if curves do not cross and by Two-Stage procedure if curves cross).

	ER positive	ER negative	PR positive	PR negative	HER2 positive	HER2 negative
α_DAEwt : α_noDAE	0.0999	0.1985	0.144	0.2515	0.9639	0.3756
α_DAEwt : α_DAEmut	0.2539	0.4142	0.6898	0.439	0.7251	0.3173
α_noDAE : α_DAEmut	0.7765	0.6347	0.8818	0.6401	0.4385	0.3125

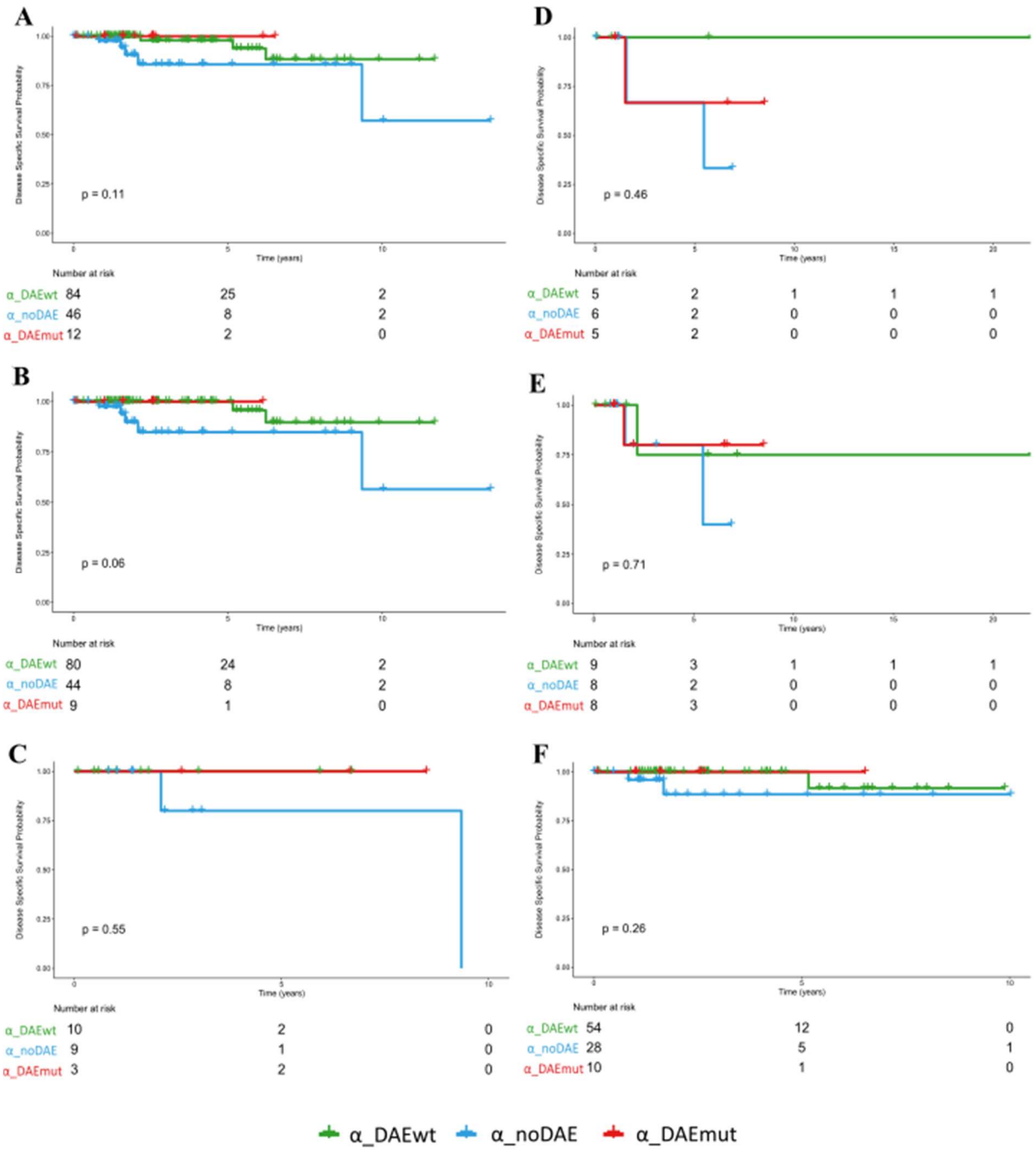


Figure H.7: Disease specific survival of *PIK3CA's* α_DAE groups in TCGA set according to ER, PR and HER2 statuses. Disease specific survival in (A) ER positive group, (B) PR positive group, (C) HER2 positive group, (D) ER negative group, (E) PR negative group and (F) HER2 negative group.

Table H.4: P-values of pairwise comparison of disease specific survival of PIK3CA's α_DAE groups in TCGA set (p-values calculated by log-rank test if curves do not cross and by Two-Stage procedure if curves cross).

	ER positive	ER negative	PR positive	PR negative	HER2 positive	HER2 negative
α_DAEwt : α_noDAE	0.1876	0.198	0.0239	0.4187	0.4385	0.2206
α_DAEwt : α_DAEmut	0.238	0.4142	0.8065	0.7591	—	0.839
α_noDAE : α_DAEmut	0.281	0.6326	0.3285	0.5637	0.4385	0.4088

Annex I – Multivariate Cox Model Regression

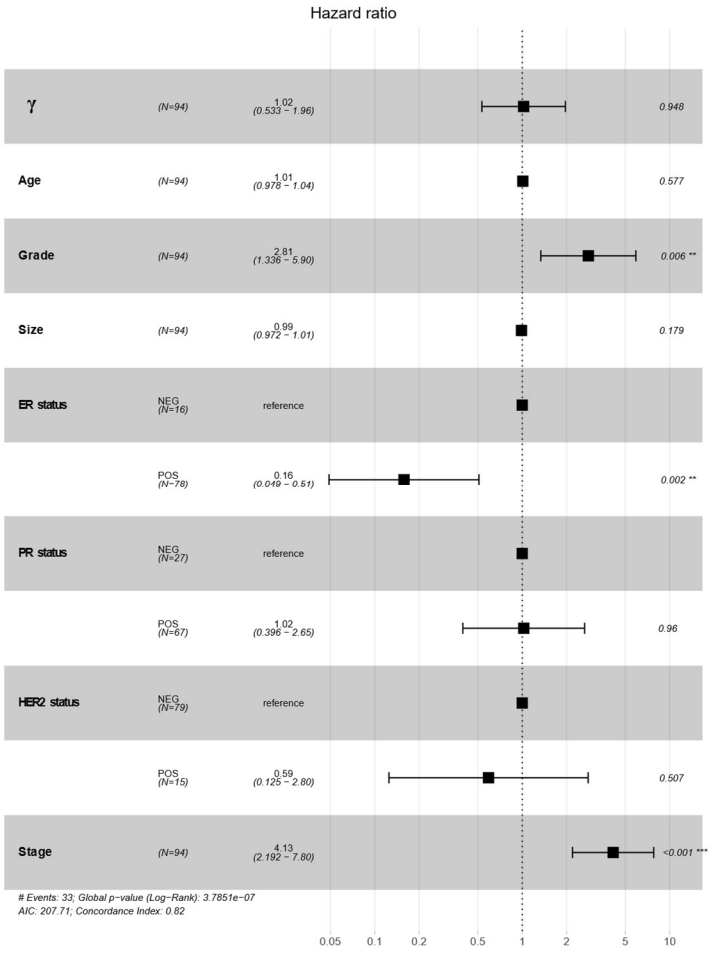


Figure I.1: Forest plot from multivariate Cox regression model showing the association of *PIK3CA's* γ ratios and overall survival, adjusted for age, grade, ER, PR and HER2 statuses and tumour stage.

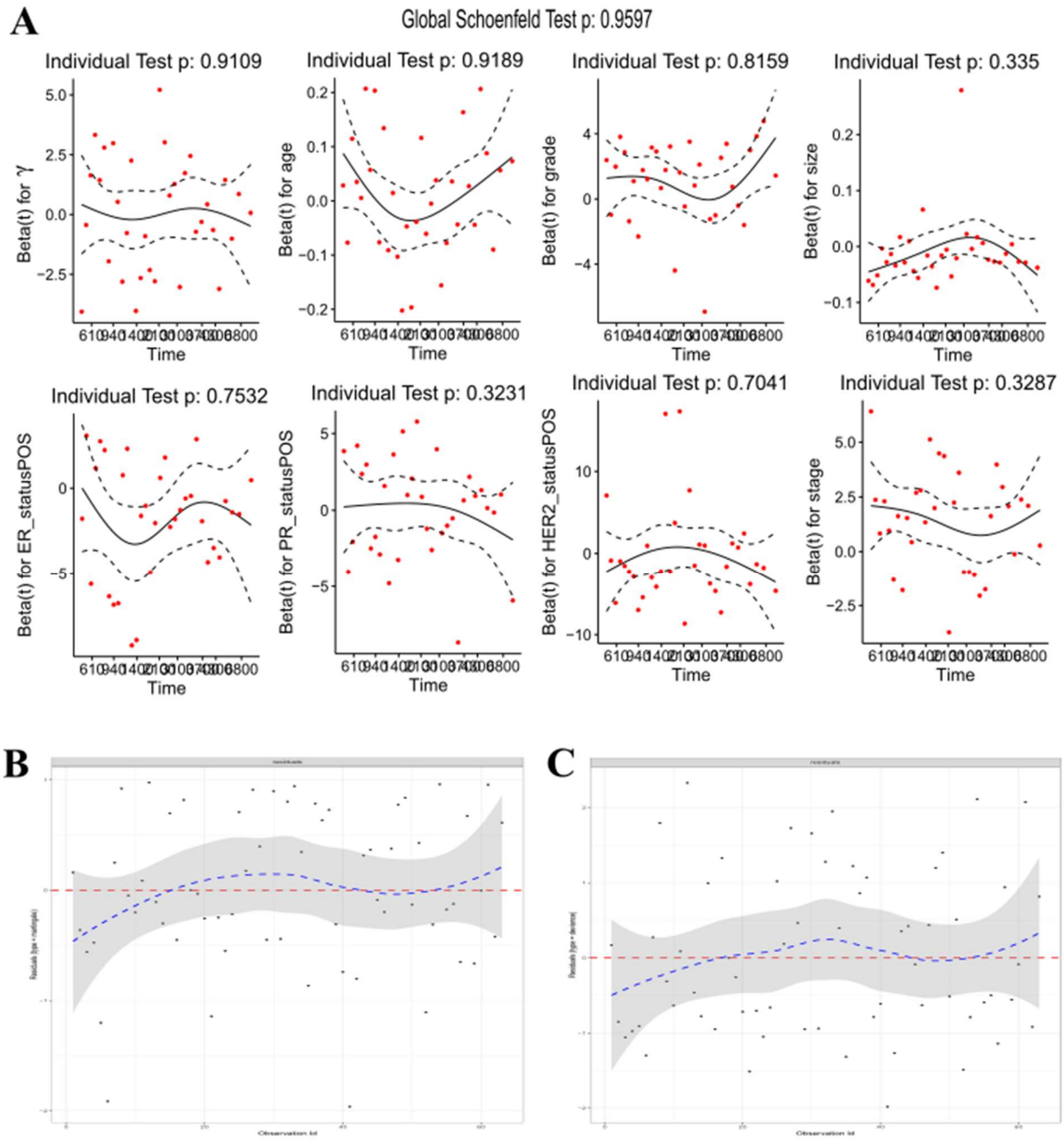


Figure I.2: Diagnostics for the fit of the Cox Model in Figure I.1. (A) Schoenfeld residual, (B) Martingale residuals an (C) Deviance residual.

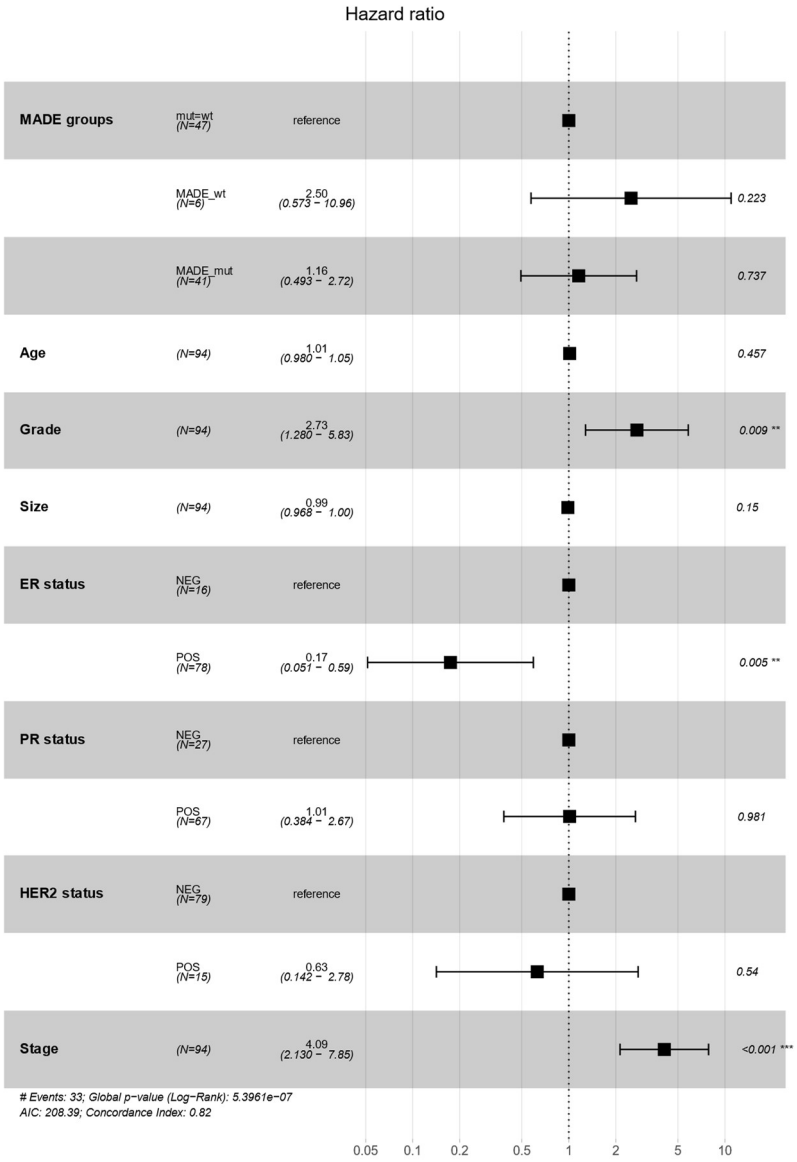


Figure I.3: Forest plot from multivariate Cox regression model showing the association of *PIK3CA's* MADE groups and overall survival, adjusted for age, grade, ER, PR and HER2 statuses and tumour stage.

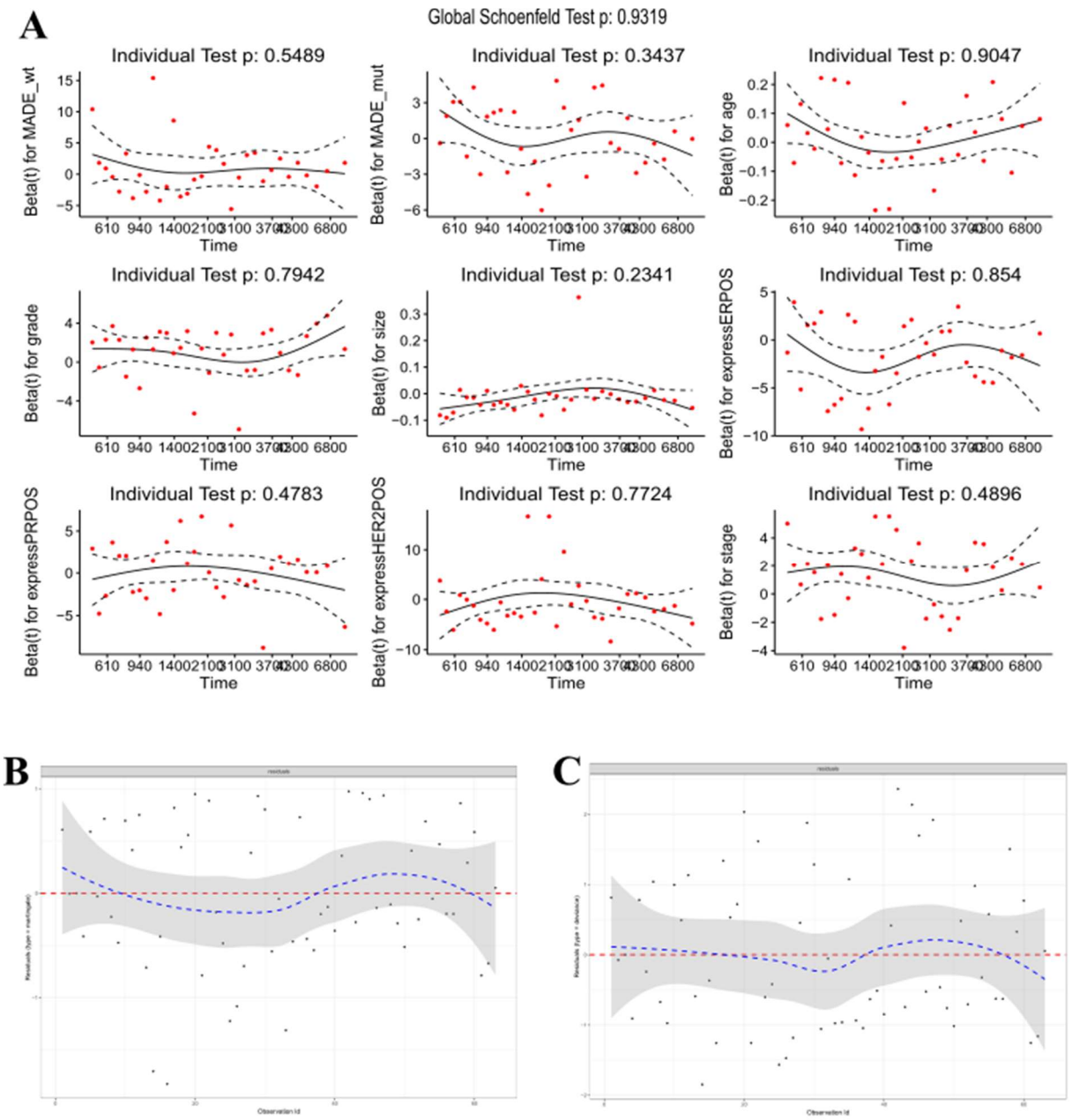


Figure I.4: Diagnostics for the fit of the Cox Model in Figure I.3. (A) Schoenfeld residual, (B) Martingale residuals an (C) Deviance residual.

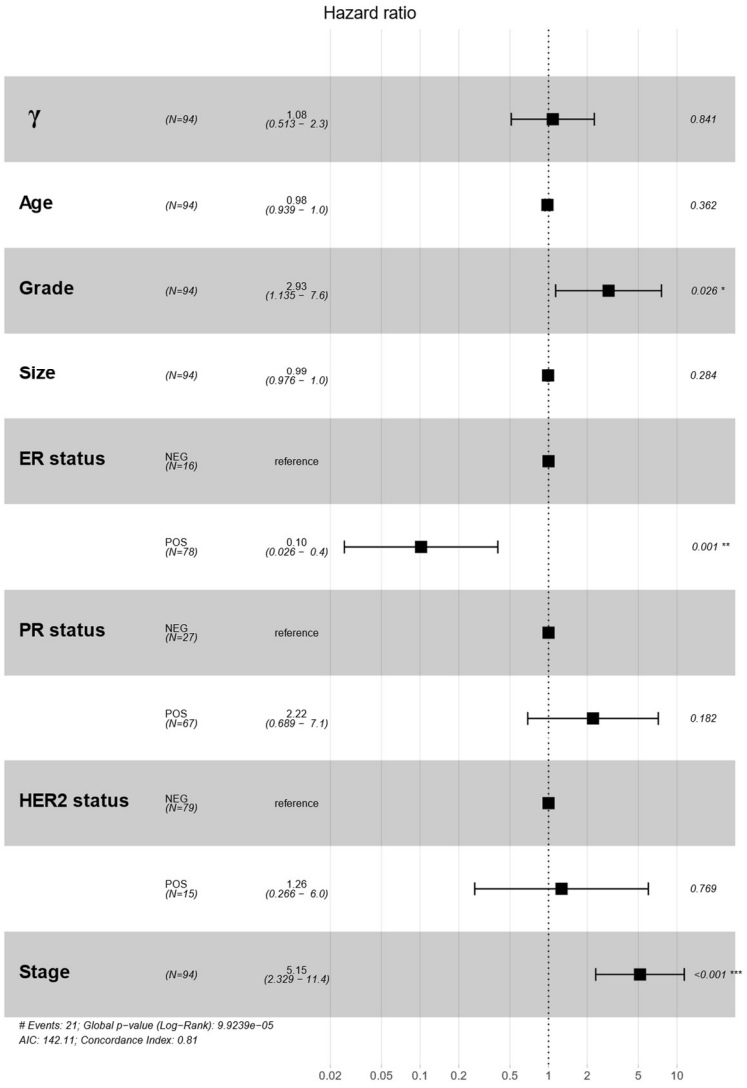


Figure I.5: Forest plot from multivariate Cox regression model showing the association of *PIK3CA's* γ ratios and disease specific survival, adjusted for age, grade, ER, PR and HER2 statuses and tumour stage.

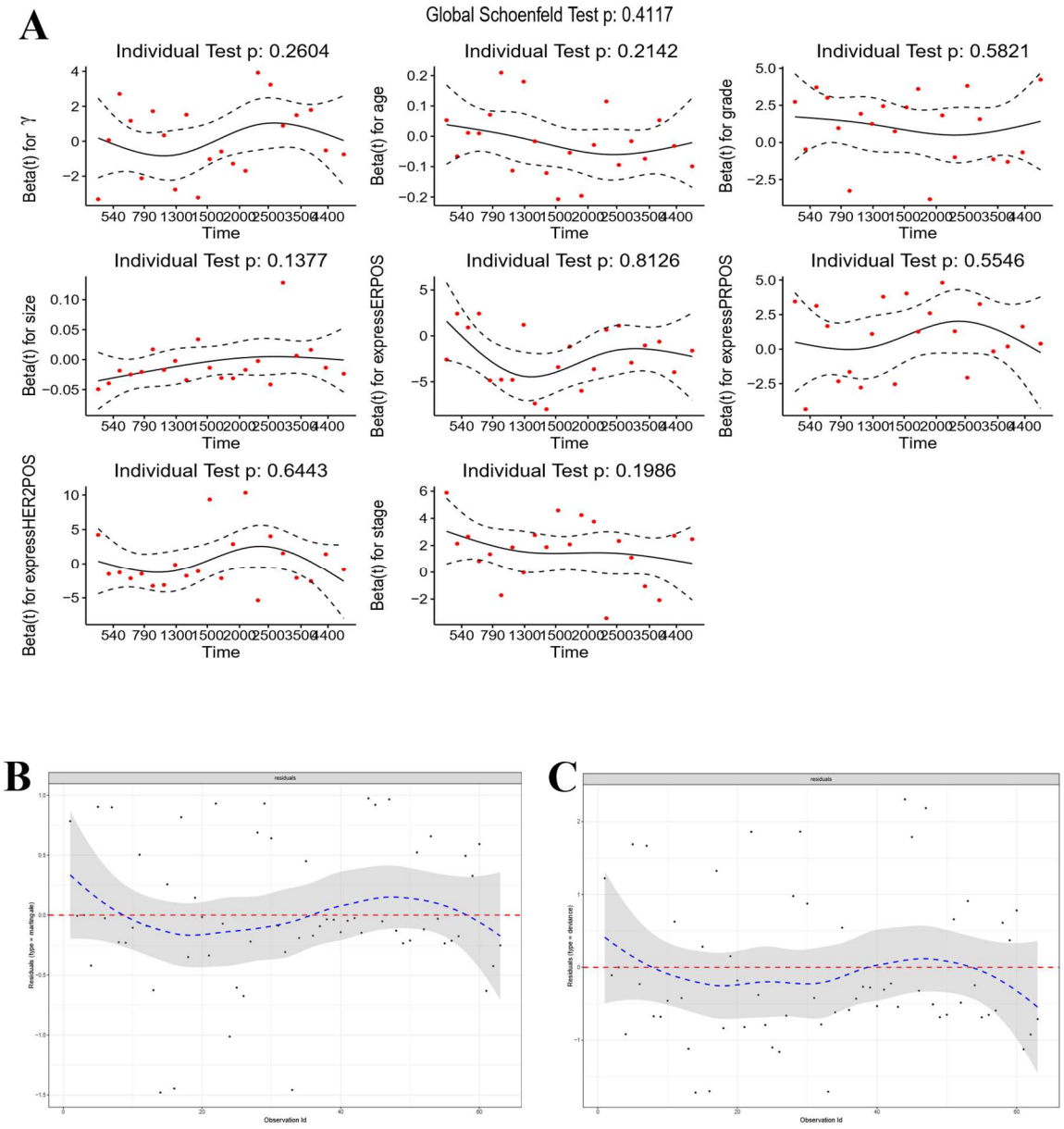


Figure I.6: Diagnostics for the fit of the Cox Model in Figure I.5. (A) Schoenfeld residual, (B) Martingale residuals an (C) Deviance residual.

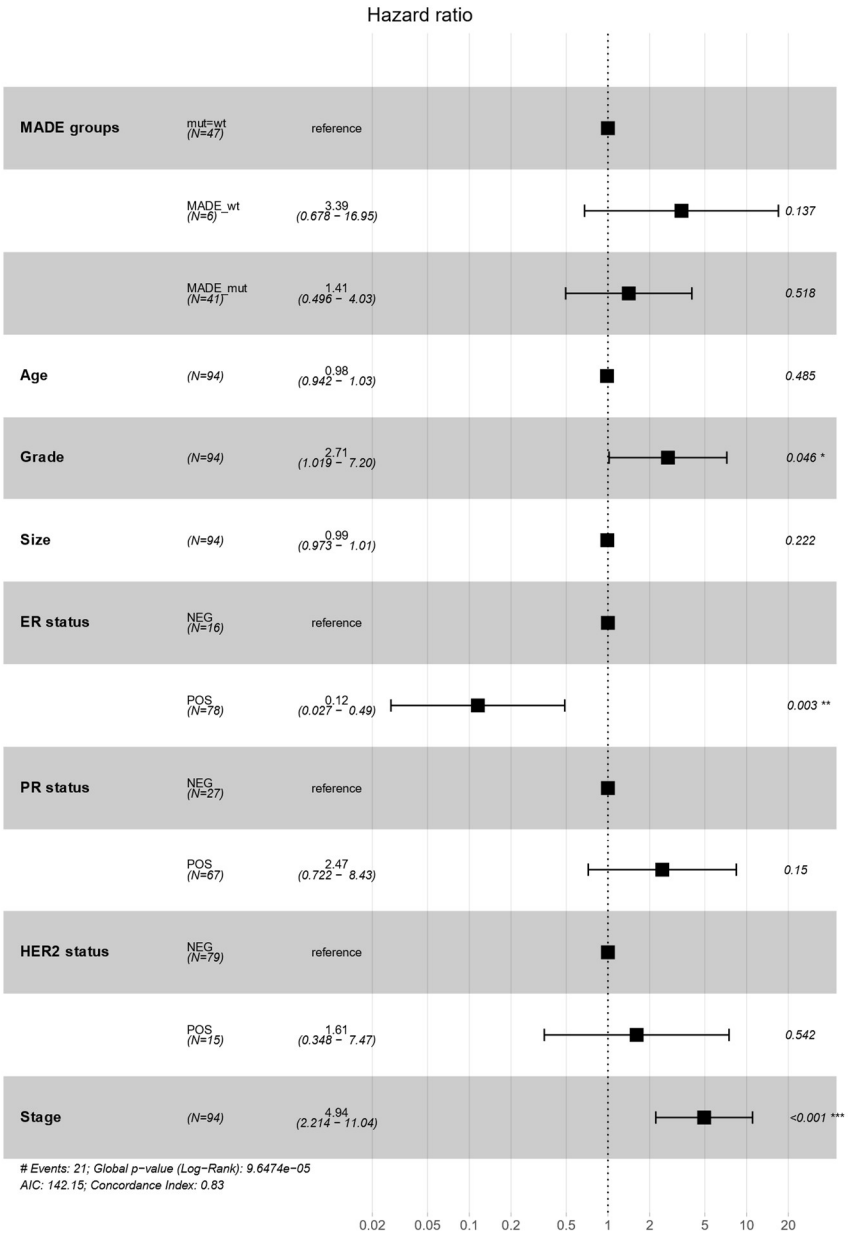


Figure I.7: Forest plot from multivariate Cox regression model showing the association of *PIK3CA's* MADE groups and disease specific survival, adjusted for age, grade, ER, PR and HER2 statuses and tumour stage.

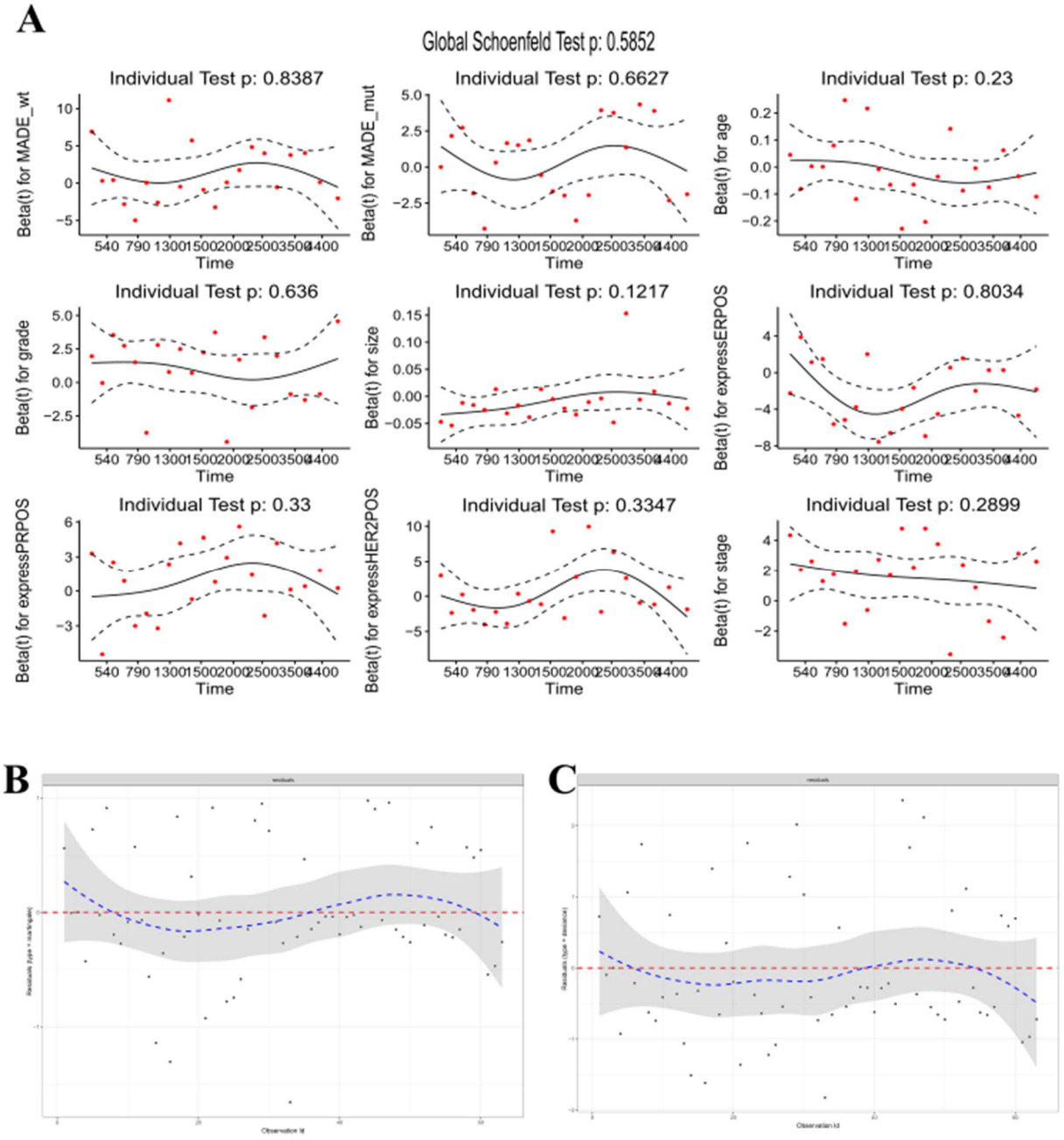


Figure I.8: Diagnostics for the fit of the Cox Model in Figure I.7. (A) Schoenfeld residual, (B) Martingale residuals an (C) Deviance residual.

Annex J

Table J.1: Statistical analysis between clinicopathological characteristics and *PIK3CA's* α and γ ratios for METABRIC set

		Shapiro-Wilk test	Mann-Whitney U test*	Kruskal- Wallis test*	Levene test	Whelch test*	Student t-test*	ANOVA*
α	ER	6.23 e-09	0.025					
	PR		0.007					
	HER2		0.018					
	Grade			0.767				
	Stage			1				
	iC			0.136				
	PAM50			0.192				
	Age group			1				
	Size group			1				
	Lymph node group			1				
γ	ER	0.02	0.129					
	PR		0.039					
	HER2		0.025					
	Grade			1				
	Stage			1				
	iC			1				
	PAM50			0.132				
	Age group			1				
	Size group			1				
	Lymph node group			1				

* p-values are adjusted with Bonferroni's correction

Table J.2: Statistical analysis between clinicopathological characteristics and *PIK3CA's* α and γ ratios for TCGA set

		Shapiro-Wilk test	Mann-Whitney U test*	Kruskal- Wallis test*	Levene test	Whelch test*	Student t-test*	ANOVA*
	ER		0.139					
	PR		0.045					
	HER2		1					
α	Stage	4.17 e-07		1				
	PAM50			0.002				
	Age group			0.03				
	Lymph node group			1				
	ER				0.855		0.009	
	PR				0.985		0.002	
	HER2				0.59		0.103	
γ	Stage	0.712			0.3306			1
	PAM50				0.9054			0.1872
	Age group				0.388			1
	Lymph node group				0.6203			1

* p-values are adjusted with Bonferroni's correction

Annex K

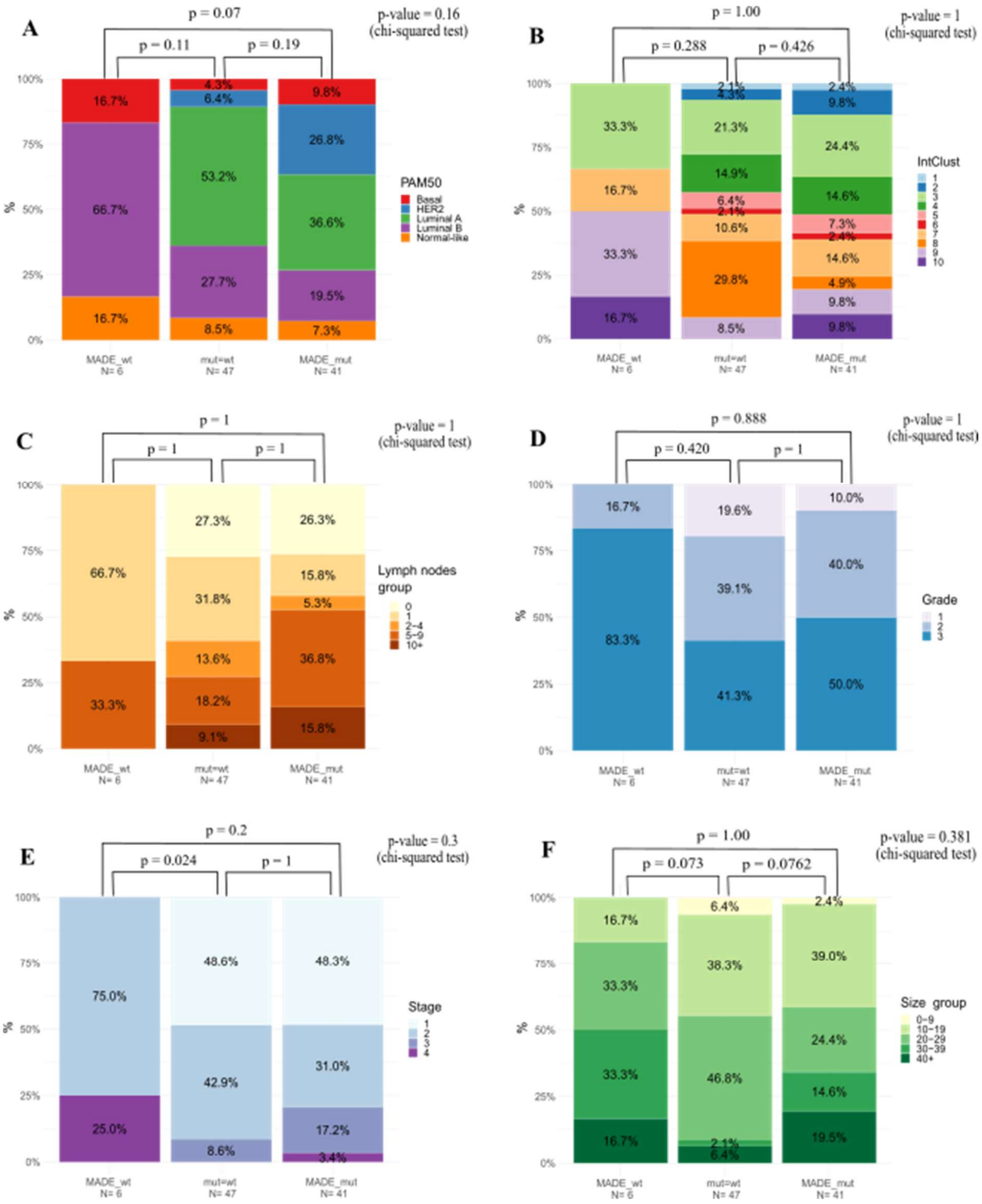


Figure K.1: Correlation of *PIK3CA's* MADE groups in METABRIC set and clinicopathological characteristics: (A) PAM50 classification; (B) Integrative Cluster Classification; (C) Lymph Node group; (D) Histological grade; (E) Tumour stage; (F) Tumour size. (IntClust: Integrative Cluster, PAM50: Prediction Analysis of Microarray 50)

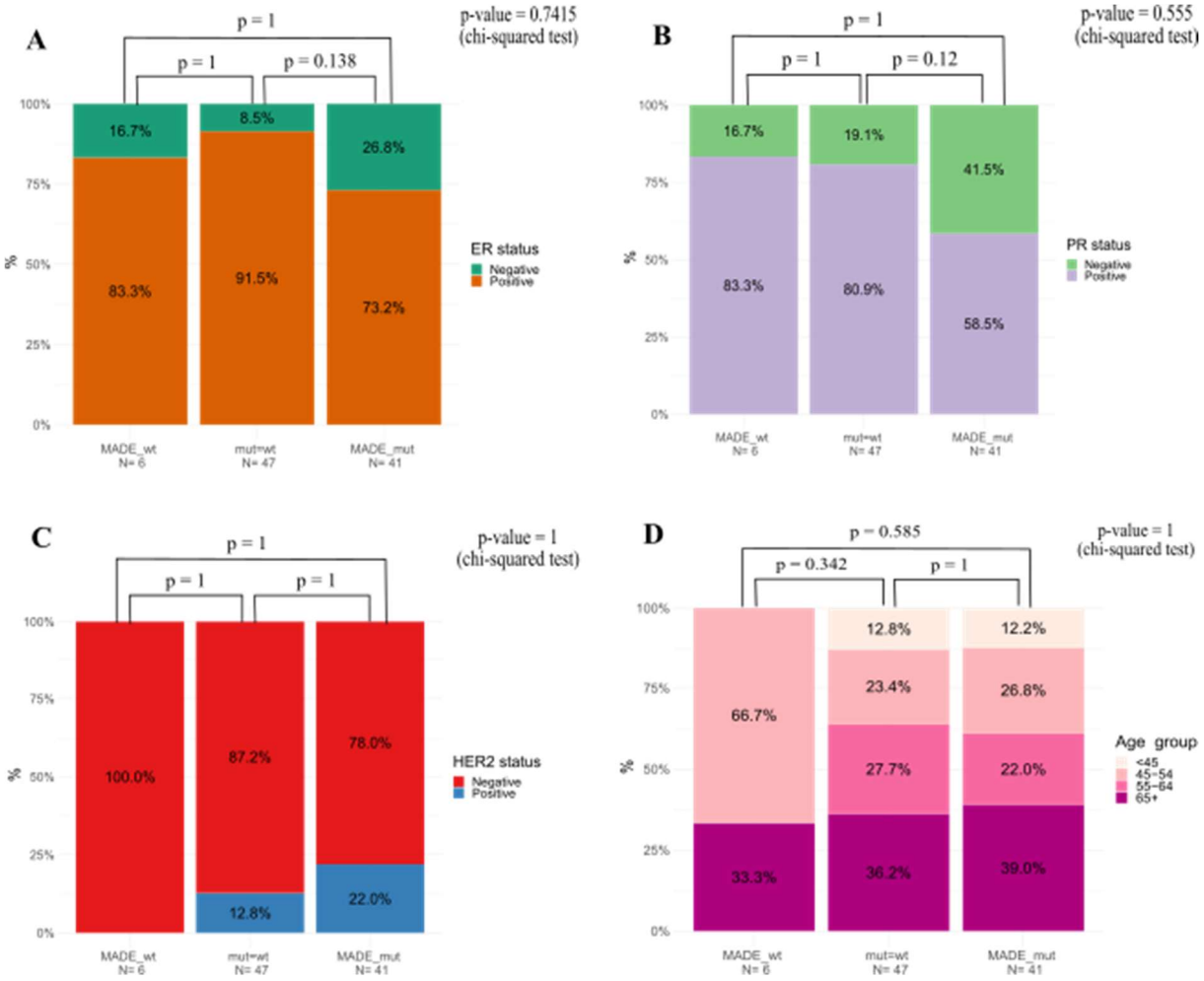
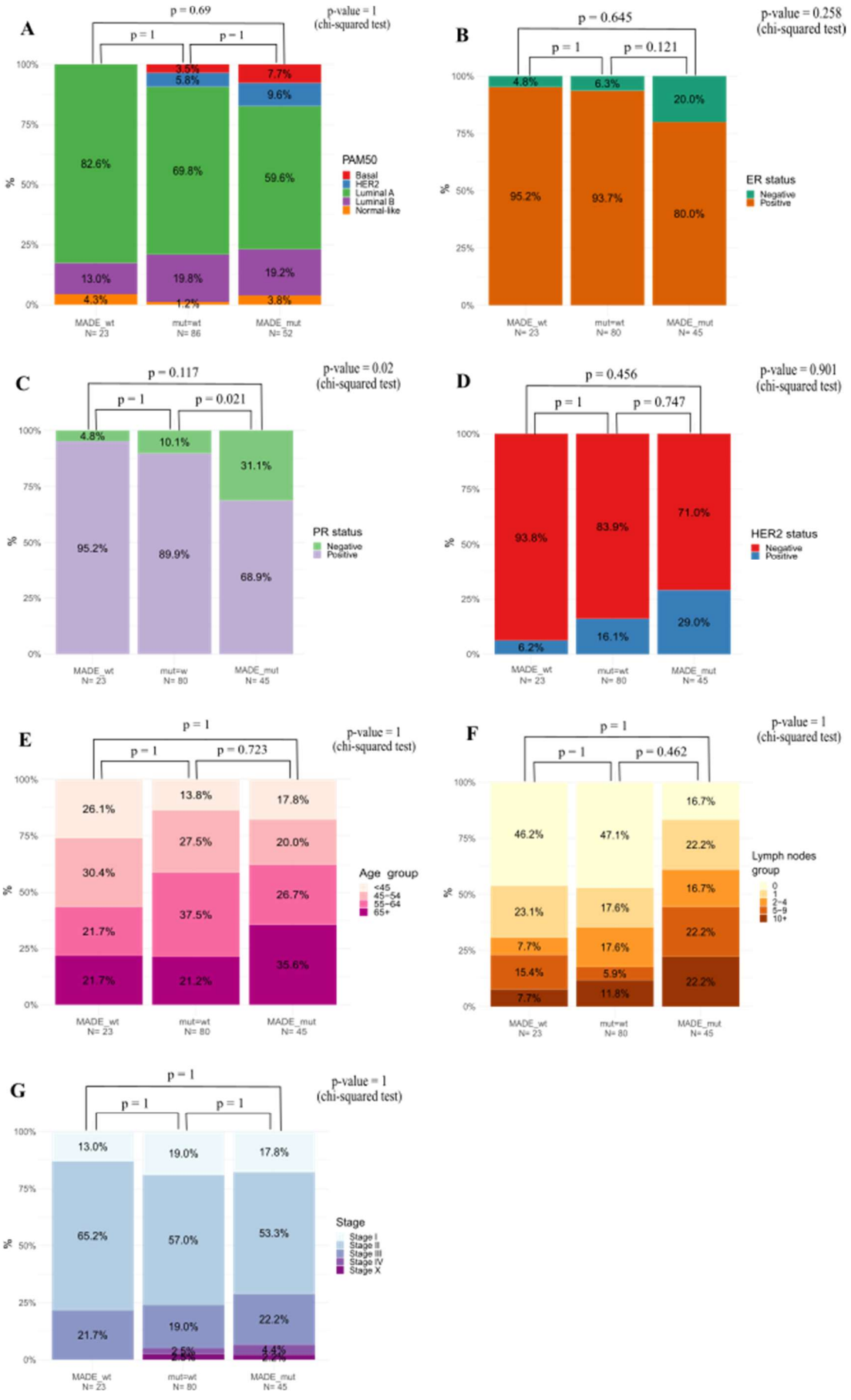


Figure K.2: Correlation of *PIK3CA's* MADE groups in METABRIC set and clinicopathological characteristics: (A) ER status; (B) PR status; (C) HER2 status; (D) Patient age.


A
p = 0.69
p = 1
p = 1
p-value = 1
(chi-squared test)
100%
75%
50%
25%
0%
%
82.6%
13.0%
4.3%
69.8%
19.8%
1.2%
3.5%
5.8%
59.6%
19.2%
3.8%
7.7%
9.6%
PAM50
Basal
HER2
Luminal A
Luminal B
Normal-like
MADE_wt
N= 23
mut=wt
N= 86
MADE_mut
N= 52
B
p = 0.645
p = 1
p = 0.121
p-value = 0.258
(chi-squared test)
4.8%
95.2%
6.3%
93.7%
20.0%
80.0%
ER status
Negative
Positive
MADE_wt
N= 23
mut=wt
N= 80
MADE_mut
N= 45
C
p = 0.117
p = 1
p = 0.021
p-value = 0.02
(chi-squared test)
4.8%
95.2%
10.1%
89.9%
31.1%
68.9%
PR status
Negative
Positive
MADE_wt
N= 23
mut=w
N= 80
MADE_mut
N= 45
D
p = 0.456
p = 1
p = 0.747
p-value = 0.901
(chi-squared test)
93.8%
6.2%
83.9%
16.1%
71.0%
29.0%
HER2 status
Negative
Positive
MADE_wt
N= 23
mut=wt
N= 80
MADE_mut
N= 45
E
p = 1
p = 1
p = 0.723
p-value = 1
(chi-squared test)
26.1%
30.4%
21.7%
21.7%
13.8%
27.5%
37.5%
21.2%
17.8%
20.0%
26.7%
35.6%
Age group
<45
45–54
55–64
65+
MADE_wt
N= 23
mut=wt
N= 80
MADE_mut
N= 45
F
p = 1
p = 1
p = 0.462
p-value = 1
(chi-squared test)
46.2%
23.1%
7.7%
15.4%
7.7%
47.1%
17.6%
17.6%
5.9%
11.8%
16.7%
22.2%
16.7%
22.2%
22.2%
Lymph nodes
group
0
1
2–4
5–9
10+
MADE_wt
N= 23
mut=wt
N= 80
MADE_mut
N= 45
G
p = 1
p = 1
p = 1
p-value = 1
(chi-squared test)
13.0%
65.2%
21.7%
19.0%
57.0%
19.0%
17.8%
53.3%
22.2%
Stage
Stage I
Stage II
Stage III
Stage IV
Stage X
MADE_wt
N= 23
mut=wt
N= 80
MADE_mut
N= 45

On previous page Figure K.3: Correlation of *PIK3CA's* MADE groups in TCGA set and clinicopathological characteristics: (A) PAM50 classification; (B) ER status; (C) PR status; (D) HER2 status; (E) Patients Age; (F) Lymph node group; (G) Tumour stage. (PAM50: Prediction Analysis of Microarray 50)

Annex L

Table L.1: Summary Statistic of rs2699887 TT, CT and CC groups Survival Analysis in METABRIC set (Median, LCL and UCL in days; LCL: lower confidence level; UCL: uper confidence level; N: number of people of each group).

		Overall Survival					Disease Specific Survival			
		N	Events	Median	0.95LCL	0.95UCL	Events	Median	0.95LCL	0.95UCL
All	TT	8	5	3617	839	NA	4	3617	839	NA
	CT	29	16	5280	2582	NA	10	NA	5280	NA
	CC	47	23	3348	3348	NA	10	NA	NA	NA
ER positive	TT	6	3	3950	3617	NA	2	NA	3617	NA
	CT	23	10	6750	5280	NA	7	NA	5280	NA
	CC	40	18	4341	3447	NA	6	NA	NA	NA
ER negative	TT	2	2	889	839	NA	2	889	839	NA
	CT	6	6	908	606	NA	3	1530	744	NA
	CC	7	5	1374	962	NA	4	1762	985	NA
PR positive	TT	6	3	3950	3617	NA	2	NA	3617	NA
	CT	18	6	NA	5280	NA	4	NA	NA	NA
	CC	34	16	4341	3348	NA	7	NA	NA	NA
PR negative	TT	2	2	889	839	NA	2	889	839	NA
	CT	11	10	1530	744	NA	6	1799	744	NA
	CC	13	7	2697	1296	NA	3	NA	2149	NA
HER2 positive	TT	0	—	—	—	—	—	—	—	—
	CT	8	6	2315	1330	NA	4	2582	1530	NA
	CC	6	2	NA	2149	NA	2	NA	2149	NA
HER2 negative	TT	8	5	3617	839	NA	4	3617	839	NA
	CT	21	10	6750	4101	NA	6	NA	5280	NA
	CC	41	21	3681	3318	NA	8	NA	NA	NA

Table L.2: Summary Statistic of rs2699887 TT, CT and CC groups Survival Analysis in METABRIC set divided in MADE and mut=wt groups (Median, LCL and UCL in days; LCL: lower confidence level; UCL: uper confidence level; N: number of people of each group).

		Overall Survival								
		N	**Events**	**Median**	**0.95LCL**	**0.95UCL**	**Events**	**Median**	**0.95LCL**	**0.95UCL**
MADE	TT	6	4	3617	939	NA	3	3617	939	NA
	CT	12	8	2056	1071	NA	5	NA	1530	NA
	CC	25	13	3447	2149	NA	9	NA	2907	NA
mut=wt	TT	2	1	730	730	NA	1	730	730	NA
	CT	17	8	6750	4138	NA	5	NA	5280	NA
	CC	22	10	4341	3660	NA	1	NA	NA	NA

Table L.3: Summary Statistic of rs2699887 TT/CT and CC groups Survival Analysis in METABRIC set (Median, LCL and UCL in days; LCL: lower confidence level; UCL: uper confidence level; N: number of people of each group).

		Overall Survival					Disease Specific Survival			
		N	Events	Median	0.95LCL	0.95UCL	Events	Median	0.95LCL	0.95UCL
All	CC	47	23	4277	3348	NA	10	NA	NA	NA
	TT/CT	37	21	4138	2582	NA	14	NA	3617	NA
ER positive	CC	40	18	4341	3447	NA	6	NA	NA	NA
	TT/CT	29	13	6750	4101	NA	9	NA	5280	NA
ER negative	CC	7	5	1374	962	NA	4	1762	985	NA
	TT/CT	8	8	889	744	NA	5	939	744	NA
PR positive	CC	34	16	4341	3348	NA	7	NA	NA	NA
	TT/CT	24	9	6750	4101	NA	6	NA	NA	NA
PR negative	CC	13	7	2697	1296	NA	3	NA	2149	NA
	TT/CT	13	12	1071	744	NA	8	1530	839	NA
HER2 positive	CC	6	2	NA	2149	NA	2	NA	2149	NA
	TT/CT	8	6	2315	1530	NA	4	2582	1530	NA
HER2 negative	CC	41	21	3681	3318	NA	8	NA	NA	NA
	TT/CT	29	15	6750	3617	NA	10	NA	5280	NA

Annex M – *PIK3CA's* Kaplan-Meier Analysis based on rs2699887 genotype

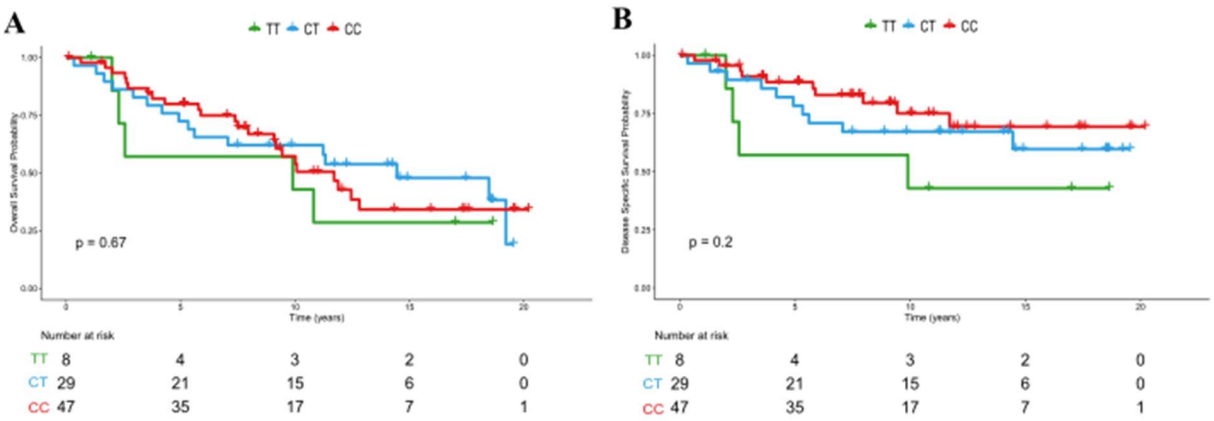


Figure M.1: Kaplan-Meier analysis of overall survival and disease specific survival of rs2699887 genotype in the METABRIC data set. (A) Overall Survival of METABRIC set divided in TT, CT and CC groups. (B) Disease Specific Survival of METABRIC set divided in TT, CT and CC groups.

Table M.1: P-values of pairwise comparison of overall and disease specific survivals of rs2699887 genotype in METABRIC set (p-values calculated by log-rank test if curves do not cross and by Two-Stage procedure if curves cross).

	OS	DSS
TT : CT	0.351	0.837
TT : CC	0.401	0.5629
CT : CC	0.251	0.669

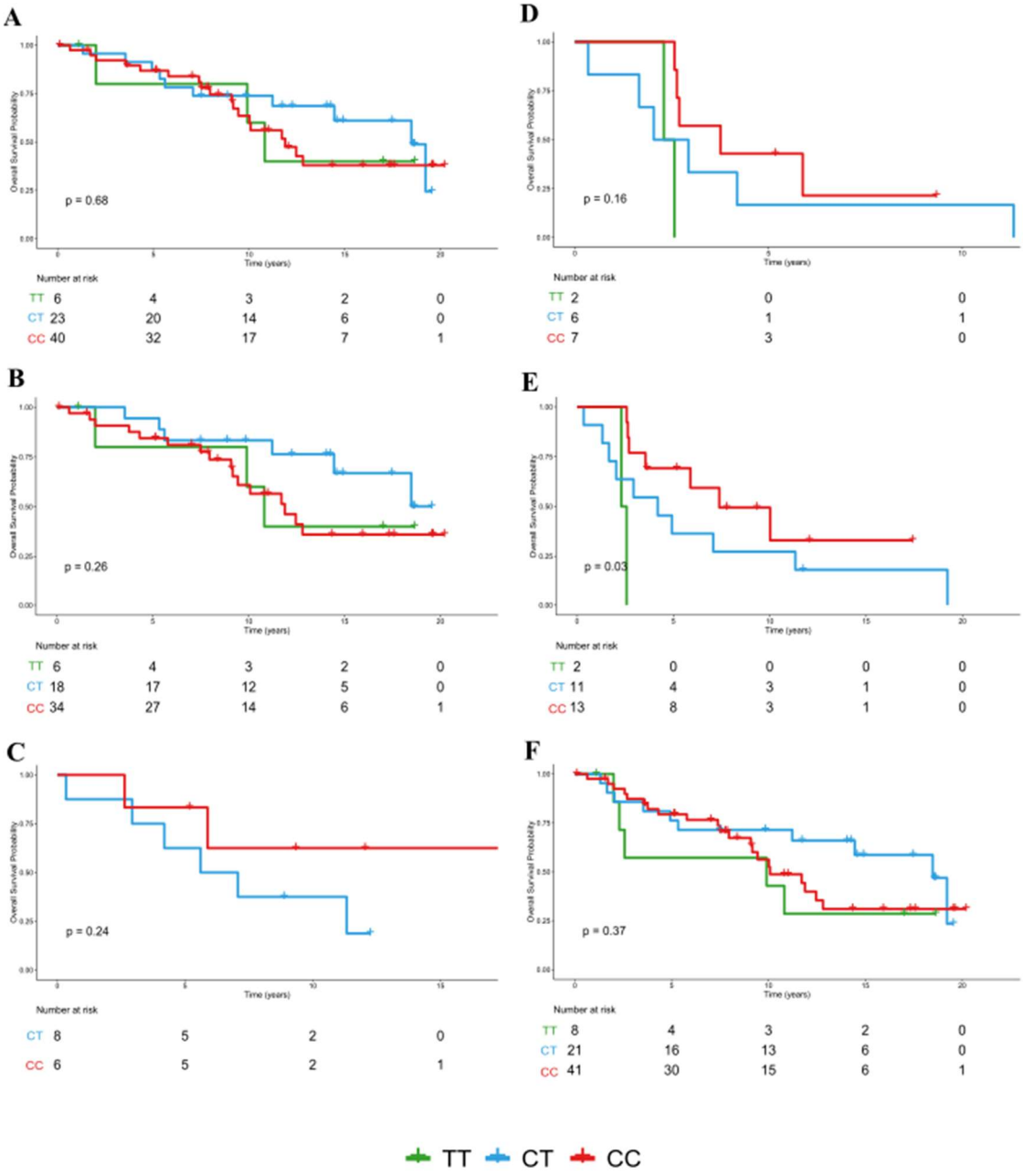


Figure M.2: Kaplan-Meier analysis of overall survival of rs2699887 genotype groups in the NETABRIC data set according to ER, PR and HER2 statuses**.** Overall survival in (A) ER positive group, (B) PR positive group, (C) HER2 positive group, (D) ER negative group, (E) PR negative group and (F) HER2 negative group.

Table M.2: P-values of pairwise comparison of overall and disease specific survivals of rs2699887 genotype in METABRIC set according to ER, PR and HER2 statuses. (p-values calculated by log-rank test if curves do not cross and by Two-Stage procedure if curves cross).

	ER positive	ER negative	PR positive	PR negative	HER2 positive	HER2 negative
TT : CT	0.905	0.966	0.881	0.486		0.804
TT : CC	0.943	0.002	0.966	2.57E-05		0.404
CT : CC	0.413	0.41	0.488	0.2165	0.849	0.262

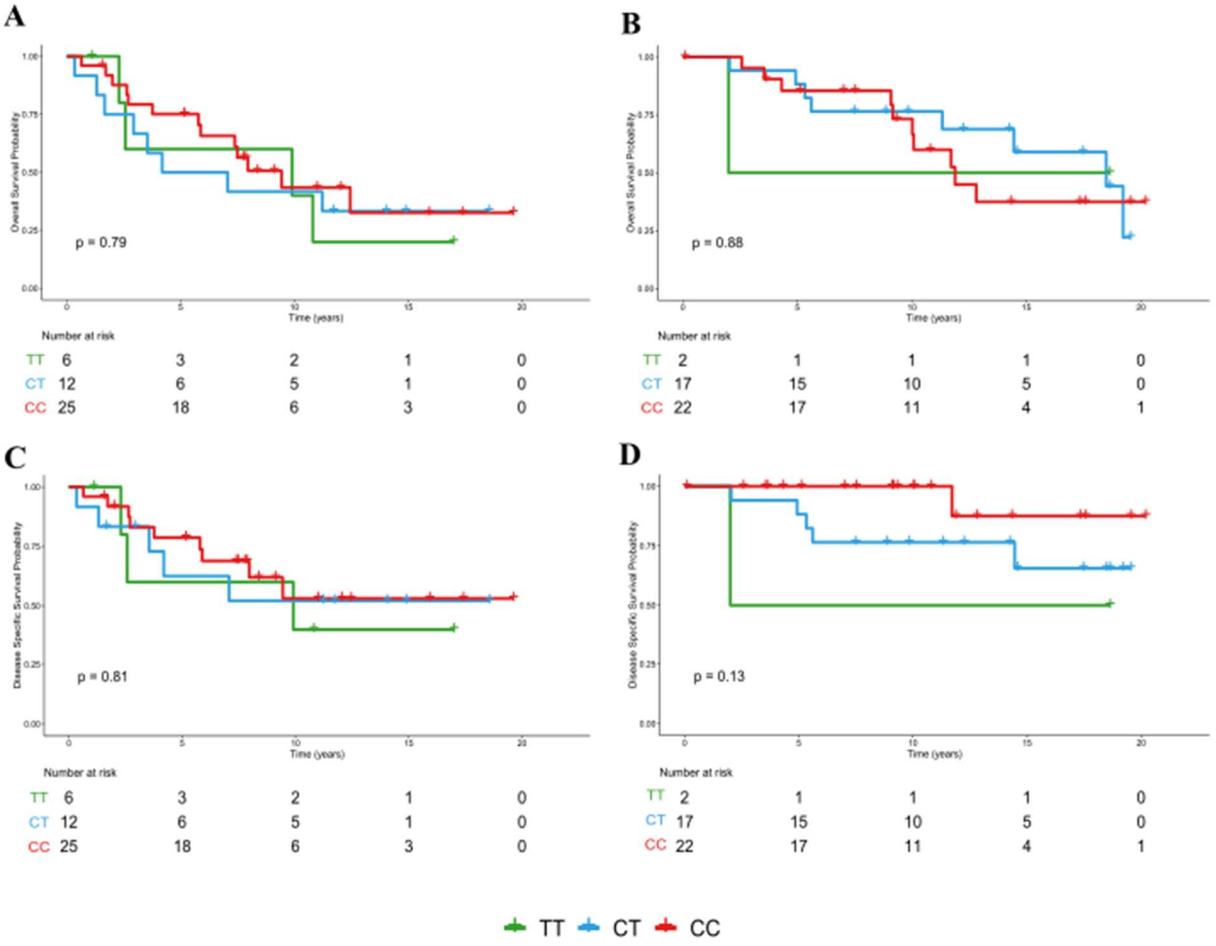


Figure M.3: Kaplan-Meier analysis of overall survival and disease specific survival of rs2699887 genotype TT, CT and CC groups in the METABRIC data set**.** (A) Overall Survival of METABRIC set MADE group divided in TT, CT and CC groups. (B) Overall Survival of METABRIC set mut=wt group divided in TT, CT and CC groups. (C) Disease Specific Survival of METABRIC set MADE group divided in TT, CT and CC groups. (D) Disease Specific Survival of METABRIC set mut=wt group divided in TT, CT and CC groups.

Table M.3: P-values of pairwise comparison of overall and disease specific survivals of rs2699887 genotype in METABRIC set MADE and mut=wt groups. (p-values calculated by log-rank test if curves do not cross and by Two-Stage procedure if curves cross; OS: Overall Survival. DSS: Disease Specific Survival).

	MADE		mut=wt	
	OS	DSS	OS	DSS
TT : CT	0.97	0.96	0.55	0.42
TT : CC	0.6	0.6	0.326	0.6
CT : CC	0.666	0.7	0.134	0.1

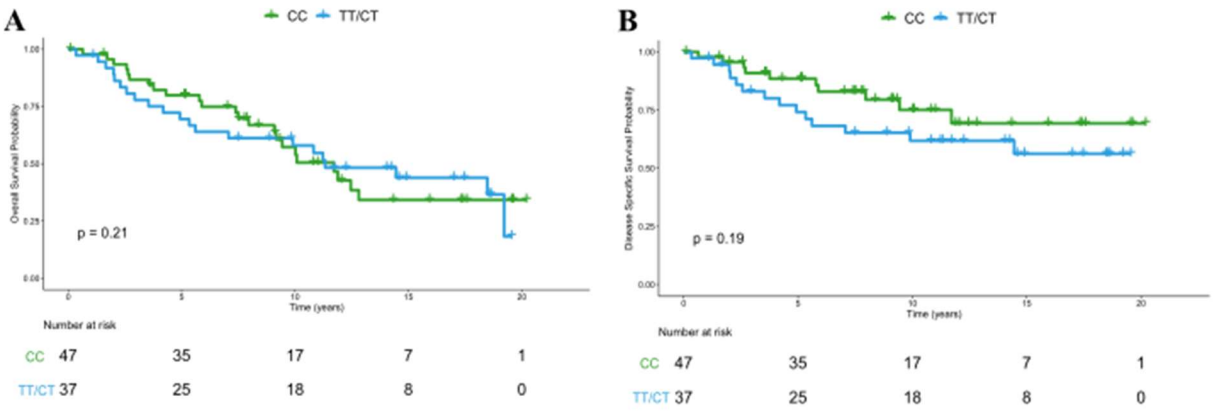


Figure M.4: Kaplan-Meier analysis of overall survival and disease specific survival of rs2699887 genotype groups in the NETABRIC data set. (A) Overall Survival of METABRIC set divided in TT/CT and CC groups. (B) Disease Specific Survival of METABRIC set divided in TT/CT and CC groups.

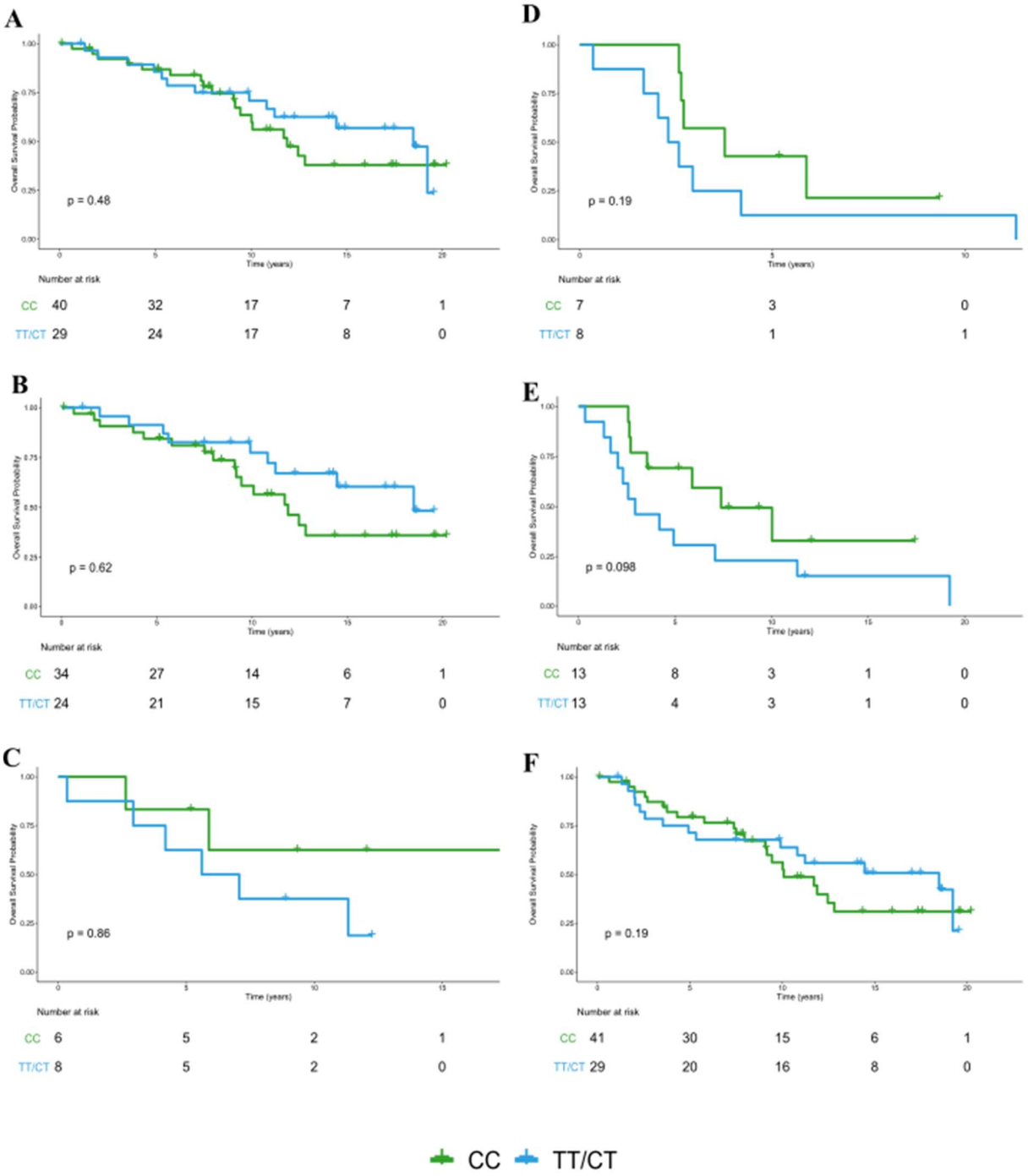


Figure M.5: Kaplan-Meier analysis of overall survival of rs2699887 genotype CC, CT/TT groups in the METABRIC data set according to ER, PR and HER2 statuses. Overall survival in (A) ER positive group, (B) PR positive group, (C) HER2 positive group, (D) ER negative group, (E) PR negative group and (F) HER2 negative group.

Annex N

Table N.1: Distribution of *TP53's* mutation impact on protein by MADE group

	METABRIC set			TCGA set		
	MADE_wt	mut=wt	MADE_mut	MADE_wt	mut=wt	MADE_mut
Likely Oncogenic	3	5	32	14	4	86
Oncogenic	0	2	12	1	0	17
UnKnown	4	0	5	1	3	1

Table N.2: P-values of pairwise test of distribution of *TP53's* mutation impact on protein by MADE group (p-values are adjusted with Bonferroni's correction).

	METABRIC	TCGA
Likely Oncogenic: Oncogenic	1	1
Likely Oncogenic: Unknown	0.0381	7.05E-06
Oncogenic: Unknown	0.0501	0.00212

Annex O

Summary of METABRIC and TCGA set for *TP53* analysis

Factor	METABRIC set	TCGA set
Total Number	63	127
	Median (LQ, UQ)	*Median (LQ, UQ)*
Age at diagnosis (years)	56.45 (46.88, 11.83)	57 (50, 64.5)
Follow-up all cases (years)	7.02 (3.41, 11.83)	1.074 (0.35, 2.88)
Follow-up still living (years)	11.44 (9.36, 15.16)	0.95 (0.084, 2.55)
Vital status		
Alive	27 (42.85%)	100 (78.74%)
Dead	36 (57.15%)	27 (21.26%)
Disease Specific Death	19 (30.15%)	18 (14.17%)
Tumour size	31.54 (20, 38)	—
Lymph nodes positive		
Number (0, 1, 2, >3)	25, 38, 0, 0	53, 16, 6, 28
Grade		
I	0	—
II	11 (17.5%)	—
III	51 (80.9%)	—
NA	1 (1.6%)	—
ER status		
Positive	27 (42.85%)	57 (44.88%)
Negative	36 (57.15%)	65 (51.18%)
NA	0	5 (3.94%)
PR status		
Positive	23 (36.5%)	54 (42.52%)
Negative	40 (63.5%)	68 (53.54%)
NA	0	5 (3.94%)
HER2 status		
Positive	13 (20.6%)	29 (22.83%)
Negative	50 (79.4%)	80 (63%)
NA or Indeterminate	0	18 (14.17%)

Factor	METABRIC set	TCGA set
Stage		
I	10 (15.87%)	20 (15.75%)
II	31 (49.2%)	71 (56%)
III	6 (9.5%)	30 (23.6%)
IV	0	5 (3.94%)
Not reported	16 (25.43%)	1 (0.71%)
PAM50 subtype		
Basal	30 (47.6%)	49 (38.58%)
HER2	11 (17.5%)	22 (17.32%)
Luminal A	8 (12.7%)	24 (18.89%)
Luminal B	11 (17.5%)	27 (21.25%)
Normal	3 (4.7%)	4 (3.25%)
NA	0	1 (0.71%)
iCluster		
iC1	1 (1.6%)	—
iC2	6 (9.5%)	—
iC3	2 (3.17%)	—
iC4	7 (11.23%)	—
iC5	8 (12.7%)	—
iC6	6 (9.5%)	—
iC7	3 (4.7%)	—
iC8	0	—
iC9	8 (12.7%)	—
iC10	22 (34.9%)	—

Annex P

Table P.1: Summary of *TP53's* α, β and γ ratios for METABRIC and TCGA sets

	METABRIC					
	min	**1st Q**	**median**	**mean**	**3rd Q**	**max**
α	-2.3219	0.2339	1.2942	1.3273	2.3718	4.9542
β	-2.7942	-1.1961	-0.5191	-0.3689	0.4249	2.9583
γ	-1.8536	0.7905	2.0233	1.6963	2.74	4.8977

	TCGA					
	min	**1st Q**	**median**	**mean**	**3rd Q**	**max**
α	-4.0875	0.6454	1.9647	1.5459	2.9115	6.8138
β	-4.9542	-1.2695	-0.3293	-0.3776	0.444	5.7814
γ	-3.2413	0.9744	2.4109	1.9234	3.082	5.0761

Table P.2: P-values of pairwise comparison between *TP53's* α, β and γ ratios means for METABRIC and TCGA sets (p-values are adjusted with Bonferroni's correction)

	METABRIC	**TCGA**
α: β	1.70E-07	8.40E-16
α: γ	0.41	0.29
β: γ	1.10E-10	< 2E-16

Annex Q

Table Q.1: Statistical analysis between clinicopathological characteristics and *TP53's* α and γ ratios for METABRIC set

		Shapiro-Wilk test	Mann-Whitney U test*	Kruskal- Wallis test*	Levene test	Whelch test*	Student t-test*	ANOVA*
α	ER	0.6891			0.09995		0.6834	
	PR				0.01692	0.2758		
	HER2				0.4159		1	
	Grade				0.07287			0.938
	Stage				0.6198			1
	iC				0.2582			0.2345
	PAM50			1	0.01801			
	Age group				0.6636			1
	Size group				0.4757			0.868
	Lymph node group				0.7384			1
γ	ER	0.06191			0.2902		0.3126	
	PR				0.1535		1	
	HER2				0.9753		1	
	Grade				0.3417			0.938
	Stage				0.4718			1
	iC				0.4128			0.2345
	PAM50				0.7631			1
	Age group				0.3629			1
	Size group				0.9523			0.868
	Lymph node group				0.3039			1

* p-values are adjusted with Bonferroni's correction

Table Q.2: Statistical analysis between clinicopathological characteristics and *TP53's* α and γ ratios for TCGA set

		Shapiro-Wilk test	Mann-Whitney U test*	Kruskal- Wallis test*	Levene test	Whelch test*	Student t-test*	ANOVA*
	ER		0.3486					
	PR		0.1304					
	HER2		1					
α	Stage	0.0022		0.2551				
	PAM50			1				
	Age group			0.682				
	Lymph node group			1				
	ER		0.03					
	PR		0.002					
	HER2		1					
γ	Stage	0.0005		0.86				
	PAM50			1				
	Age group			1				
	Lymph node group			1				

* p-values are adjusted with Bonferroni's correction

Annex R

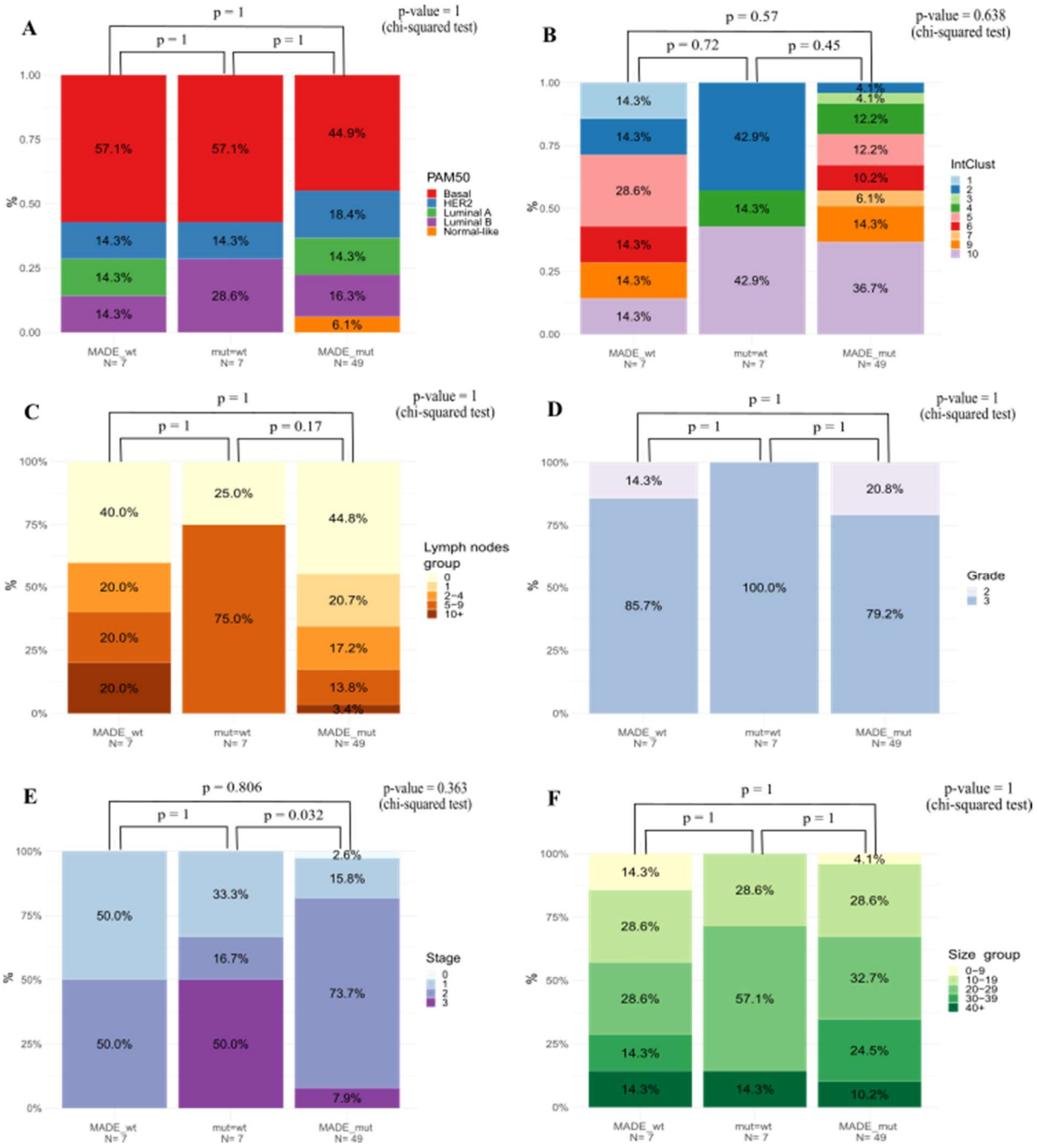


Figure R.1: Correlation of *TP53's* MADE groups in METABRIC set and clinicopathological characteristics: (A) PAM50 classification; (B) Integrative Cluster Classification; (C) Lymph Node group; (D) Histological grade; (E) Tumour stage; (F) Tumour size. (IntClust: Integrative Cluster, PAM50: Prediction Analysis of Microarray 50)

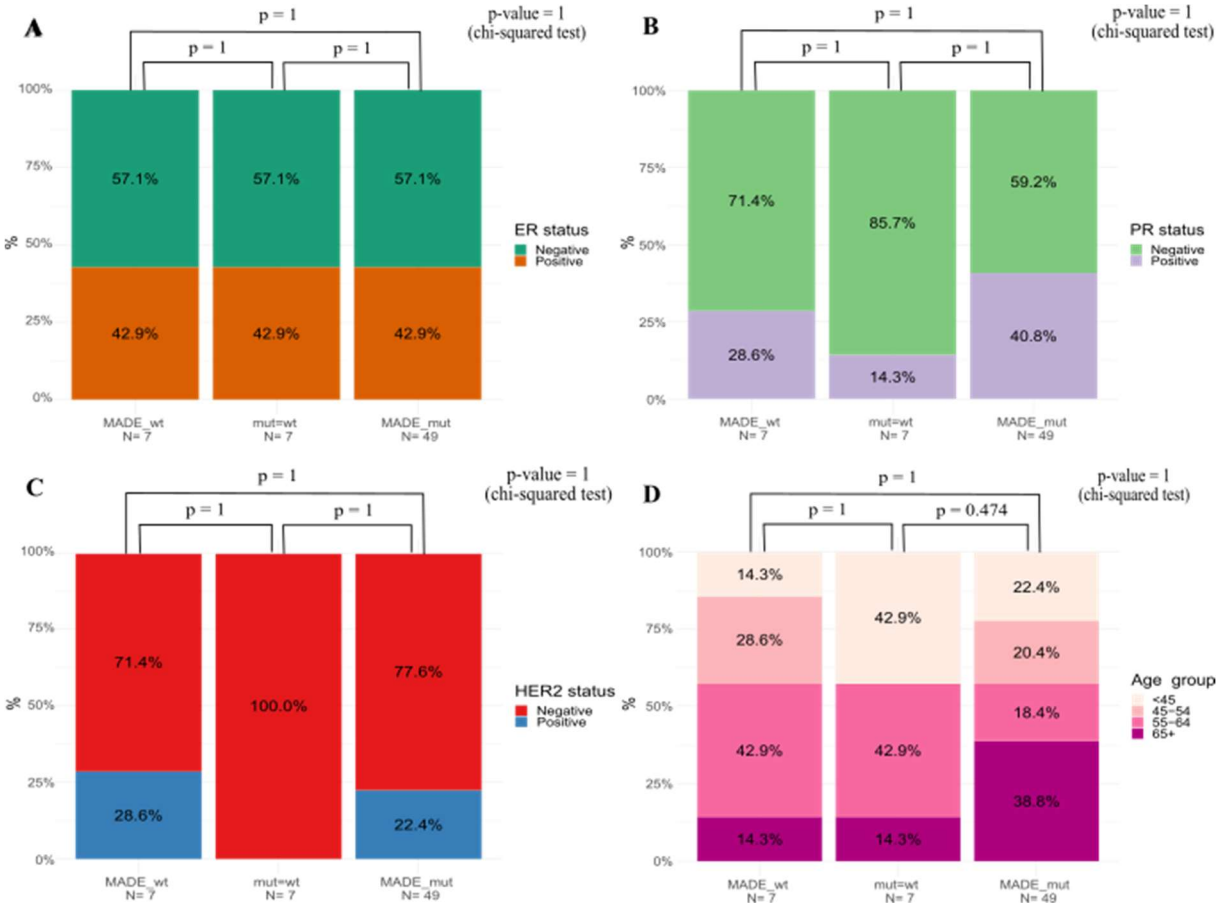
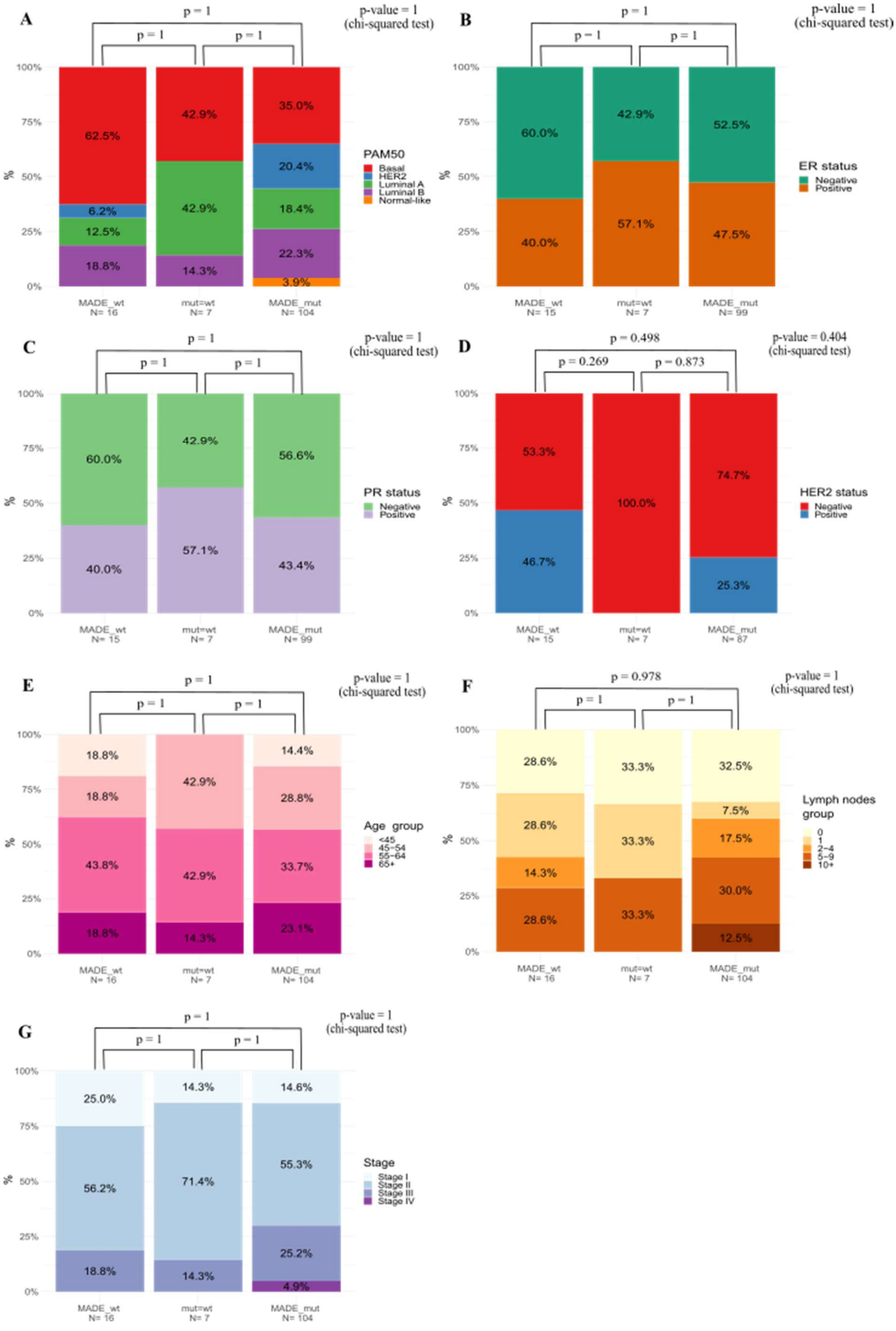


Figure R.2: Correlation of *TP53's* MADE groups in METABRIC set and clinicopathological characteristics: (A) ER status; (B) PR status; (C) HER2 status; (D) Patient age.


A
p = 1
p = 1
p = 1
p-value = 1
(chi-squared test)
100%
75%
50%
25%
0%
%
62.5%
6.2%
12.5%
18.8%
42.9%
42.9%
14.3%
35.0%
20.4%
18.4%
22.3%
3.9%
PAM50
Basal
HER2
Luminal A
Luminal B
Normal-like
MADE_wt
N= 16
mut=wt
N= 7
MADE_mut
N= 104
B
p = 1
p = 1
p = 1
p-value = 1
(chi-squared test)
60.0%
40.0%
42.9%
57.1%
52.5%
47.5%
ER status
Negative
Positive
MADE_wt
N= 15
mut=wt
N= 7
MADE_mut
N= 99
C
p = 1
p = 1
p = 1
p-value = 1
(chi-squared test)
60.0%
40.0%
42.9%
57.1%
56.6%
43.4%
PR status
Negative
Positive
MADE_wt
N= 15
mut=wt
N= 7
MADE_mut
N= 99
D
p = 0.498
p = 0.269
p = 0.873
p-value = 0.404
(chi-squared test)
53.3%
46.7%
100.0%
74.7%
25.3%
HER2 status
Negative
Positive
MADE_wt
N= 15
mut=wt
N= 7
MADE_mut
N= 87
E
p = 1
p = 1
p = 1
p-value = 1
(chi-squared test)
18.8%
18.8%
43.8%
18.8%
42.9%
42.9%
14.3%
14.4%
28.8%
33.7%
23.1%
Age_group
<45
45–54
55–64
65+
MADE_wt
N= 16
mut=wt
N= 7
MADE_mut
N= 104
F
p = 0.978
p = 1
p = 1
p-value = 1
(chi-squared test)
28.6%
28.6%
14.3%
28.6%
33.3%
33.3%
33.3%
32.5%
7.5%
17.5%
30.0%
12.5%
Lymph nodes
group
0
1
2–4
5–9
10+
MADE_wt
N= 16
mut=wt
N= 7
MADE_mut
N= 104
G
p = 1
p = 1
p = 1
p-value = 1
(chi-squared test)
25.0%
56.2%
18.8%
14.3%
71.4%
14.3%
14.6%
55.3%
25.2%
4.9%
Stage
Stage I
Stage II
Stage III
Stage IV
MADE_wt
N= 16
mut=wt
N= 7
MADE_mut
N= 104

On previous page Figure R.3: Correlation of *TP53's* MADE groups in TCGA set and clinicopathological characteristics: (A) PAM50 classification; (B) ER status; (C) PR status; (D) HER2 status; (E) Patients Age; (F) Lymph node group; (G) Tumour stage. (PAM50: Prediction Analysis of Microarray 50)

Annex S

Table S.1: Summary Statistic of *TP53's* MADE Survival Analysis in METABRIC set (Median, LCL and UCL in days; LCL: lower confidence level; UCL: uper confidence level; N: number of people of each group).

		Overall Survival					Disease Specific Survival				
		N	**Events**	**Median**	**0.95LCL**	**0.95UCL**	**N**	**Events**	**Median**	**0.95LCL**	**0.95UCL**
All	MADE	56	34	3997	1919	NA	56	15	NA	4223	NA
	mut=wt	7	4	5280	1294	NA	7	4	5280	1294	NA
ER positive	MADE	24	13	4223	3270	NA	24	5	NA	4223	NA
	mut=wt	3	2	5280	1343	NA	3	2	5280	1343	NA
ER negative	MADE	32	19	1919	1296	NA	32	10	NA	2002	NA
	mut=wt	4	2	1294	696	NA	4	2	1294	696	NA
PR positive	MADE	22	13	3997	2173	NA	22	8	4223	NA	NA
	mut=wt	1	1	5280	NA	NA	1	1	5280	3997	NA
PR negative	MADE	34	19	3121	1551	NA	34	7	NA	NA	NA
	mut=wt	6	3	1343	1294	NA	6	3	1343	1294	NA
HER2 positive	MADE	13	6	NA	NA	NA	13	3	NA	NA	NA
	mut=wt	0	—	—	—	—	0	—	—	—	—
HER2 negative	MADE	43	26	3270	1916	NA	43	12	NA	4223	NA
	mut=wt	7	4	5280	1294	NA	7	4	5280	1294	NA

Table S.2: Summary Statistic of *TP53's* MADE groups Survival Analysis in METABRIC set (Median, LCL and UCL in days; LCL: lower confidence level; UCL: uper confidence level; N: number of people of each group).

		Overall Survival					Disease Specific Survival				
		N	**Events**	**Median**	**0.95LCL**	**0.95UCL**	**N**	**Events**	**Median**	**0.95LCL**	**0.95UCL**
All	MADE_wt	7	2	NA	4138	NA	7	1	NA	NA	NA
	mut=wt	7	4	5280	1294	NA	7	4	5280	1294	NA
	MADE_mut	49	30	3270	1916	5468	49	14	NA	4223	NA
ER positive	MADE_wt	3	0	NA	NA	NA	3	0	NA	NA	NA
	mut=wt	3	2	5280	1343	NA	3	2	5280	1343	NA
	MADE_mut	21	13	3997	3121	NA	21	5	NA	4223	NA
ER negative	MADE_wt	4	2	4138	1296	NA	4	1	NA	1296	NA
	mut=wt	4	2	1294	696	NA	4	2	1294	696	NA
	MADE_mut	28	17	1916	1050	NA	28	9	NA	1374	NA
PR positive	MADE_wt	2	0	NA	NA	NA	2	0	NA	NA	NA
	mut=wt	1	1	5280	NA	NA	1	1	5280	NA	NA
	MADE_mut	20	3	3372	1374	NA	20	8	4223	3270	NA
PR negative	MADE_wt	5	2	4138	1296	NA	5	1	NA	1296	NA
	mut=wt	6	3	1343	1294	NA	6	3	1343	1294	NA
	MADE_mut	29	17	2002	1255	NA	29	6	NA	NA	NA
HER2 positive	MADE_wt	2	1	4138	NA	NA	2	0	NA	NA	NA
	mut=wt	0	—	—	—	—	0	—	—	—	—
	MADE_mut	11	5	5370	1332	NA	11	3	NA	1332	NA
HER2 negative	MADE_wt	5	1	NA	NA	NA	5	1	NA	NA	NA
	mut=wt	7	4	5280	1294	NA	7	4	5280	1294	NA
	MADE_mut	38	25	3121	1551	NA	38	11	NA	3997	NA

Table S.3: Summary Statistic of *TP53's* α_DAE groups Survival Analysis in METABRIC set (Median, LCL and UCL in days; LCL: lower confidence level; UCL: uper confidence level; N: number of people of each group).

		Overall Survival					Disease Specific Survival				
		N	**Events**	**Median**	**0.95LCL**	**0.95UCL**	**N**	**Events**	**Median**	**0.95LCL**	**0.95UCL**
All	α_DAE	52	29	4138	2173	NA	52	15	NA	5280	NA
	α_noDAE	11	7	1916	1294	NA	11	4	NA	1294	NA
ER positive	α_DAE	21	12	4223	3270	NA	21	6	NA	4223	NA
	α_noDAE	6	3	5468	1235	NA	6	1	NA	NA	NA
ER negative	α_DAE	31	17	2002	1296	NA	31	9	NA	2002	NA
	α_noDAE	5	4	1332	1294	NA	5	3	1332	1294	NA
PR positive	α_DAE	17	11	3997	3270	NA	17	7	4223	3997	NA
	α_noDAE	6	3	1332	1235	NA	6	2	NA	1332	NA
PR negative	α_DAE	35	18	4138	1551	NA	35	8	NA	NA	NA
	α_noDAE	5	4	1916	1294	NA	5	2	NA	1294	NA
HER2 positive	α_DAE	9	4	5370	4138	NA	9	1	NA	NA	NA
	α_noDAE	4	2	1332	1050	NA	4	2	1332	1050	NA
HER2 negative	α_DAE	43	25	3372	1919	NA	43	14	5280	3997	NA
	α_noDAE	7	5	1916	1235	NA	7	2	NA	1294	NA

Table S.4: Summary Statistic of *TP53's* α_DAE groups Survival Analysis in METABRIC set (Median, LCL and UCL in days; LCL: lower confidence level; UCL: uper confidence level; N: number of people of each group).

		Overall Survival					Disease Specific Survival				
		N	**Events**	**Median**	**0.95LCL**	**0.95UCL**	**N**	**Events**	**Median**	**0.95LCL**	**0.95UCL**
All	α_DAEwt	10	3	NA	2173	NA	10	1	NA	NA	NA
	α_noDAE	11	7	1916	1294	NA	11	4	NA	1294	NA
	α_DAEmut	42	26	3372	1551	5370	42	14	5280	3997	NA
ER positive	α_DAEwt	4	1	NA	2173	NA	4	0	NA	NA	NA
	α_noDAE	6	3	5468	1235	NA	6	1	NA	NA	NA
	α_DAEmut	17	11	4223	3270	NA	17	6	5280	3997	NA
ER negative	α_DAEwt	6	2	NA	4138	NA	6	1	NA	NA	NA
	α_noDAE	5	4	1332	1294	NA	5	3	1332	1294	NA
	α_DAEmut	25	15	1919	913	NA	25	8	NA	1374	NA
PR positive	α_DAEwt	3	1	2173	2173	NA	3	0	NA	NA	NA
	α_noDAE	6	3	1332	1235	NA	6	2	NA	1332	NA
	α_DAEmut	14	10	3997	3270	NA	14	7	4223	3270	NA
PR negative	α_DAEwt	7	2	NA	4138	NA	7	1	NA	NA	NA
	α_noDAE	5	4	1916	1294	NA	5	2	NA	1294	NA
	α_DAEmut	28	16	2002	1255	NA	28	7	NA	NA	NA
HER2 positive	α_DAEwt	2	1	4138	NA	NA	2	0	NA	NA	NA
	α_noDAE	4	2	1332	1050	NA	4	2	1332	1050	NA
	α_DAEmut	7	3	5370	1293	NA	7	1	NA	NA	NA
HER2 negative	α_DAEwt	8	2	NA	2173	NA	8	1	NA	NA	NA
	α_noDAE	7	3	1916	1235	NA	7	2	NA	1294	NA
	α_DAEmut	35	23	3121	1374	NA	35	13	5280	3270	NA

Annex T – *TP53's* Kaplan-Meier Survival Analysis for METABRIC set

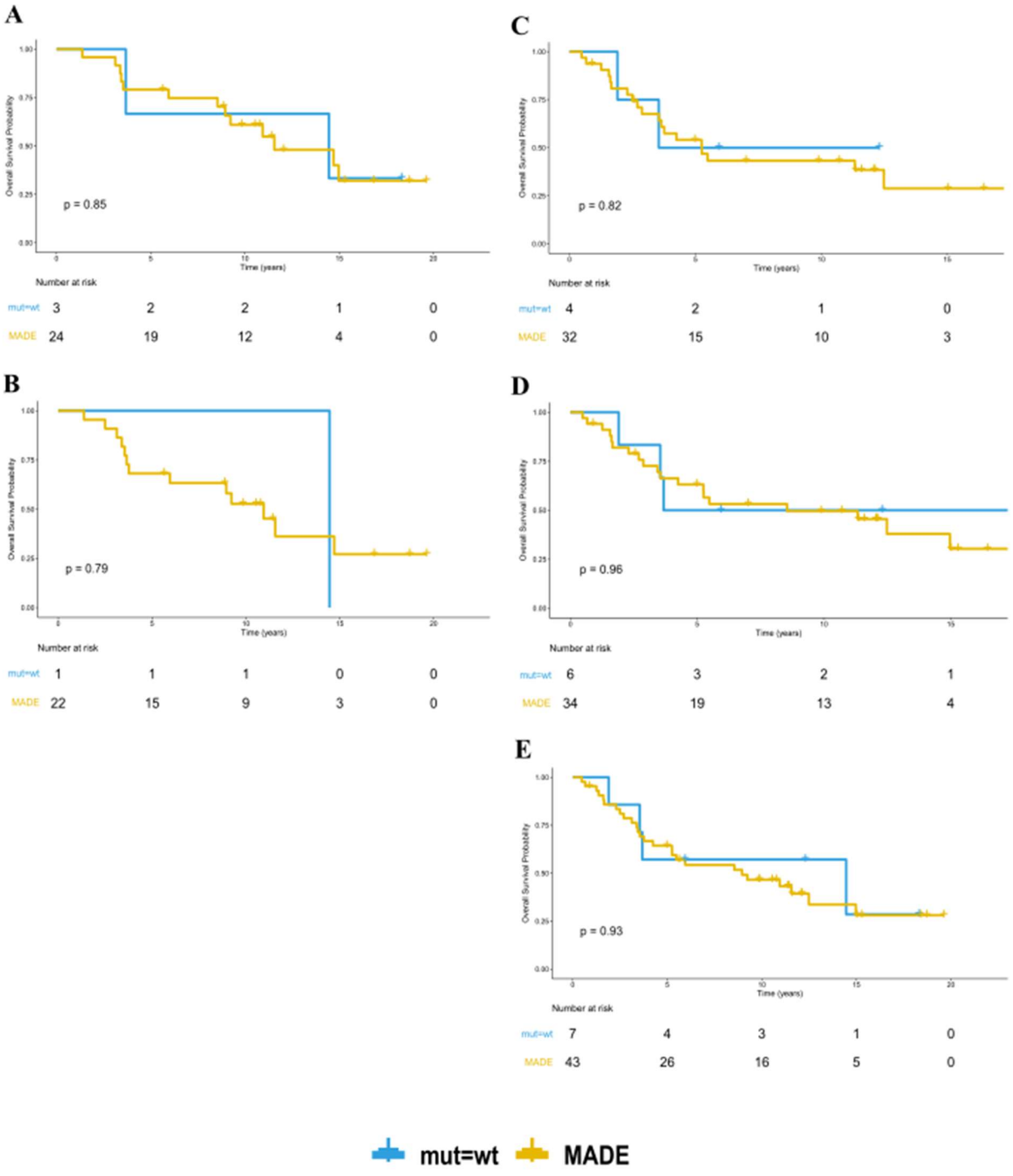


Figure T.1: Kaplan-Meier analysis of overall survival of *TP53's* MADE in the METABRIC set according to ER, PR and HER2 statuses. Overall survival in (A) ER positive group, (B) PR positive group, (C) ER negative group, (D) PR negative group and (E) HER2 negative group.

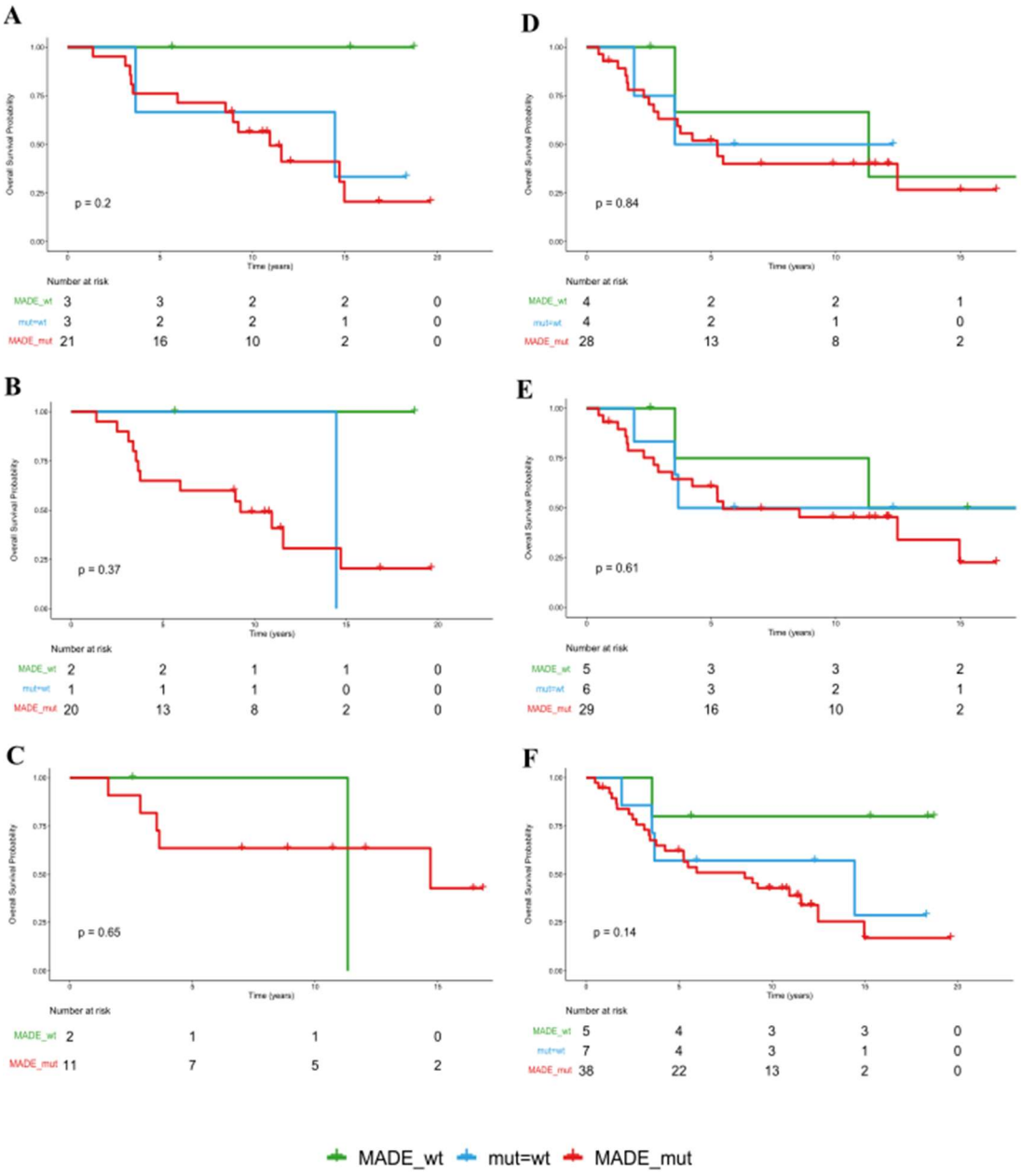


Figure T.2: Kaplan-Meier analysis of overall survival of *TP53's* MADE groups in the METABRIC set according to ER, PR and HER2 statuses. Overall survival in (A) ER positive group, (B) PR positive group, (C) HER2 positive group, (D) ER negative group, (E) PR negative group and (F) HER2 negative group.

Table T.1: P-values of pairwise comparison of overall survival of *TP53's* MADE in METABRIC set (p-values calculated by log-rank test if curves do not cross and by Two-Stage procedure if curves cross).

	ER positive	ER negative	PR positive	PR negative	HER2 positive	HER2 negative
MADE_wt : mut=wt	0.1572	0.738	0.3173	0.6998		0.2014
MADE_wt : MADE_mut	0.3008	0.3897	0.1745	0.3247	0.65	0.0578
mut=wt : MADE_mut	0.6872	0.759	0.2223	0.8834		0.8099

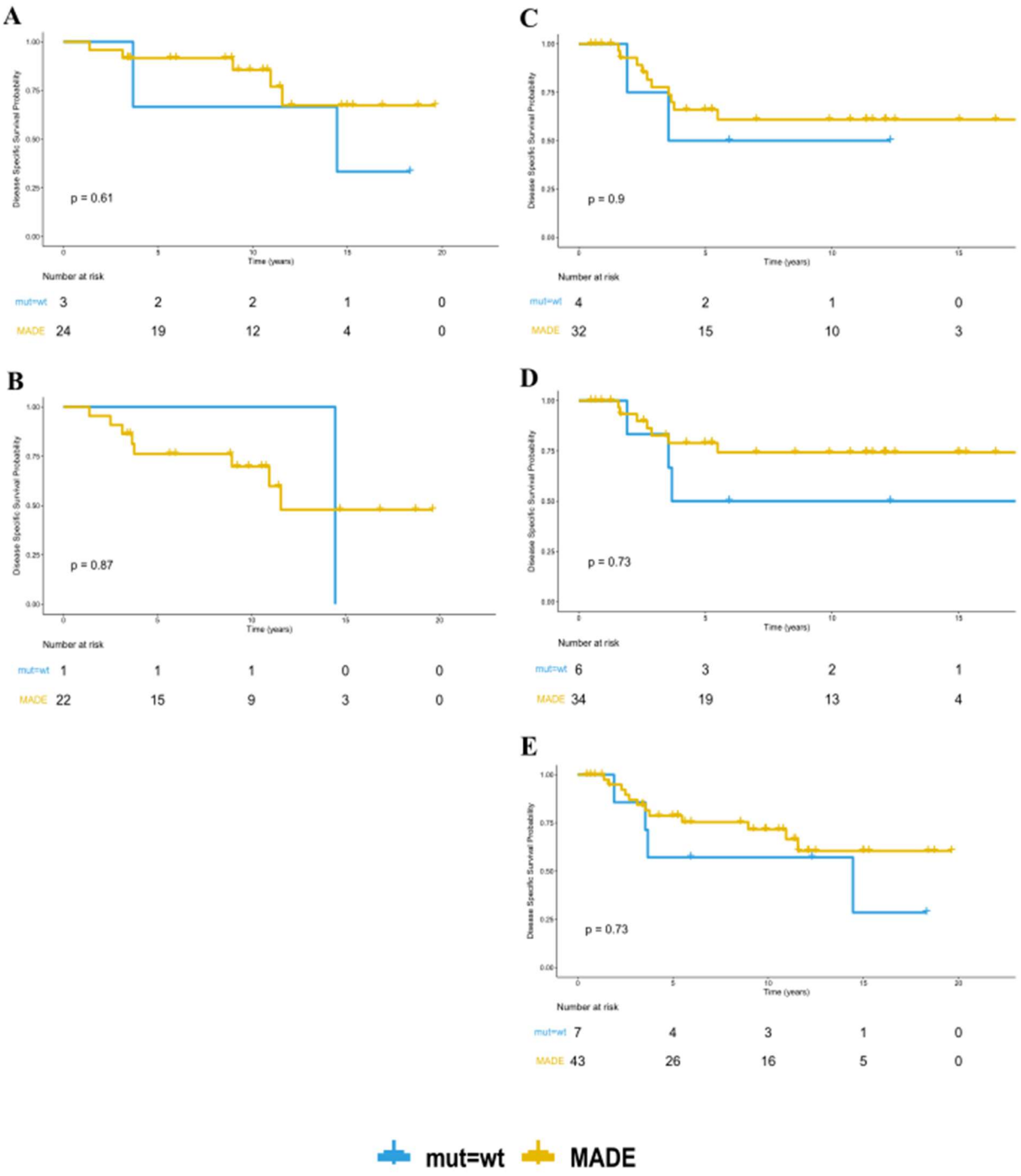


Figure T.3: Kaplan-Meier analysis of disease specific survival of *TP53's* MADE in the METABRIC set according to ER, PR and HER2 statuses. Disease specific survival in (A) ER positive group, (B) PR positive group, (C) ER negative group, (D) PR negative group and (E) HER2 negative group.

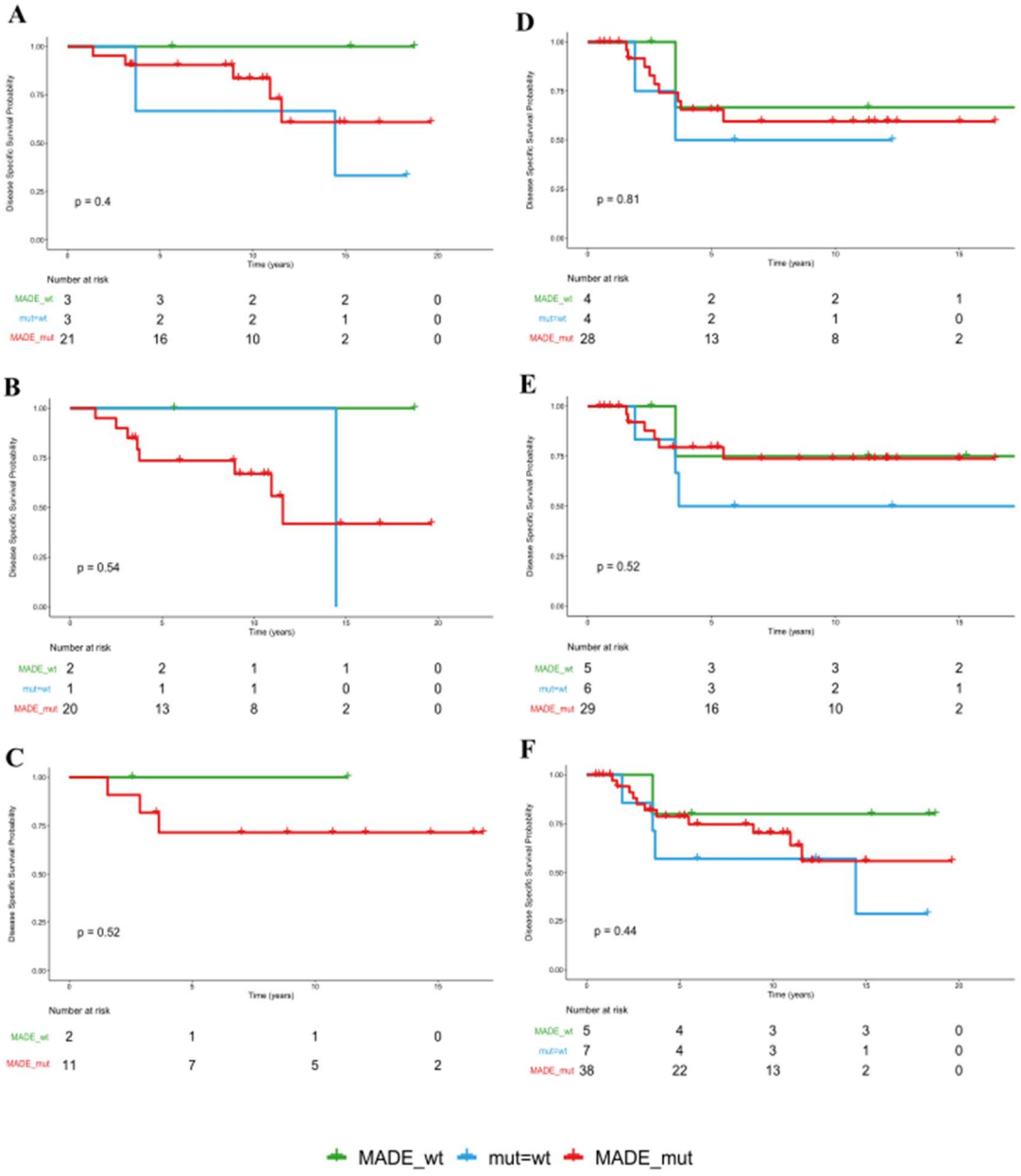


Figure T.4: Kaplan-Meier analysis of disease specific survival of *TP53's* MADE groups in the METABRIC set according to ER, PR and HER2 statuses. Disease specific survival in (A) ER positive group, (B) PR positive group, (C) HER2 positive group, (D) ER negative group, (E) PR negative group and (F) HER2 negative group.

Table T.2: P-values of pairwise comparison of disease specific survival of *TP53's* MADE in METABRIC set (p-values calculated by log-rank test if curves do not cross and by Two-Stage procedure if curves cross).

	ER positive	ER negative	PR positive	PR negative	HER2 positive	HER2 negative
MADE_wt : mut=wt	0.1387	0.5053	0.3173	0.3519		0.1815
MADE_wt : MADE_mut	0.3111	0.5747	0.6821	0.6821	0.52	0.8937
mut=wt : MADE_mut	0.8066	0.995	0.9606	0.8336		0.837

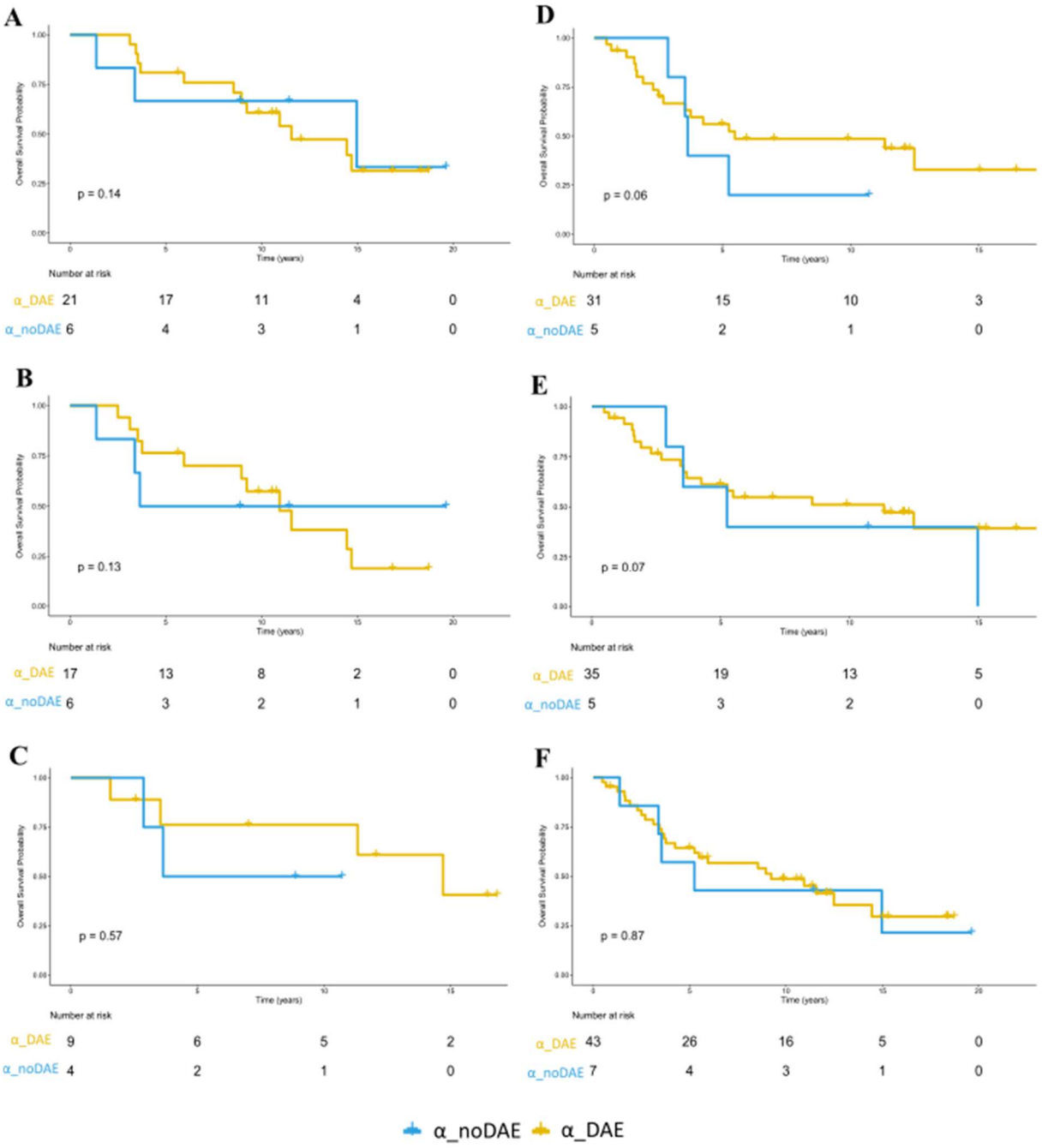


Figure T.5: Kaplan-Meier analysis of overall survival of *TP53's* α_DAE groups in the METABRIC set according to ER, PR and HER2 statuses. Overall survival in (A) ER positive group, (B) PR positive group, (C) HER2 positive group, (D) ER negative group, (E) PR negative group and (F) HER2 negative group.

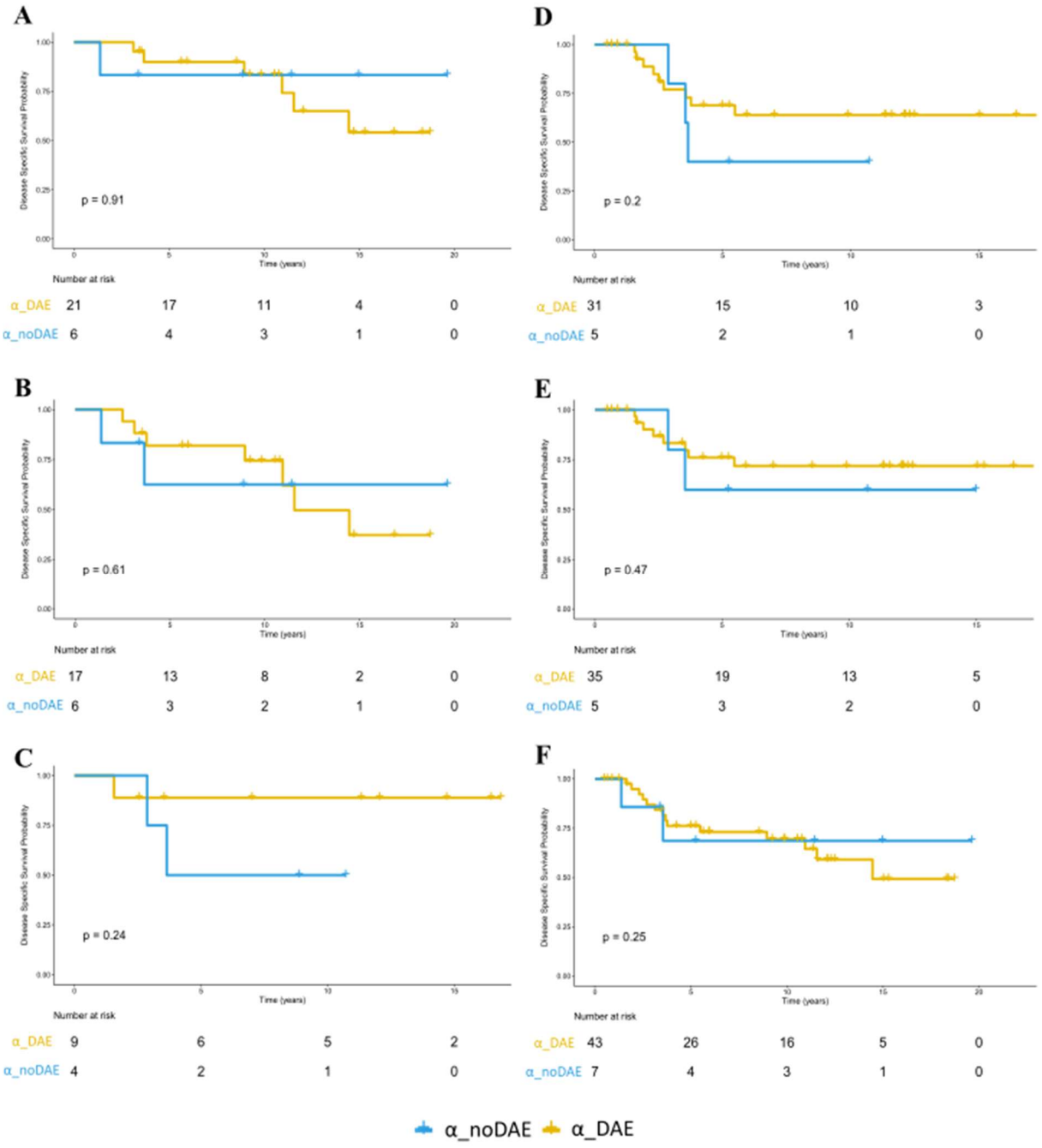


Figure T.6: Kaplan-Meier analysis of disease specific survival of *TP53's* α_DAE groups in the METABRIC set according to ER, PR and HER2 statuses. Disease specific survival in (A) ER positive group, (B) PR positive group, (C) HER2 positive group, (D) ER negative group, (E) PR negative group and (F) HER2 negative group.

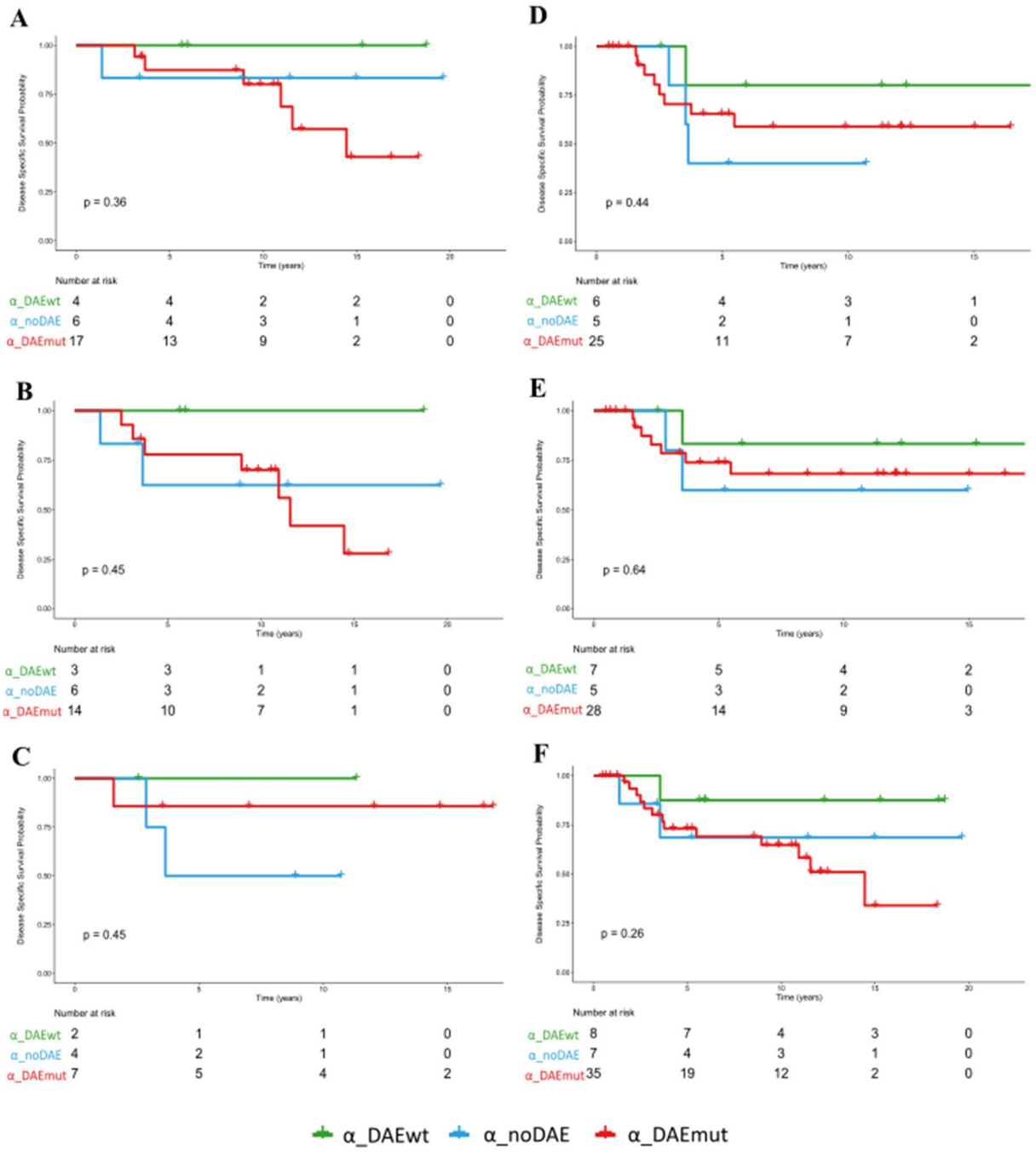


Figure T.7: Kaplan-Meier analysis of disease specific survival of *TP53's* α_DAE groups in the METABRIC set according to ER, PR and HER2 statuses. Disease specific survival in (A) ER positive group, (B) PR positive group, (C) HER2 positive group, (D) ER negative group, (E) PR negative group and (F) HER2 negative group.

Table T.3: P-values of pairwise comparison of disease specific survival of *TP53's* α_DAE groups in METABRIC set (p-values calculated by log-rank test if curves do not cross and by Two-Stage procedure if curves cross).

	ER positive	ER negative	PR positive	PR negative	HER2 positive	HER2 negative
α_DAEwt : α_noDAE	0.4142	0.1599	0.2649	0.3351	0.4452	0.3306
α_DAEwt : α_DAEmut	0.1275	0.6093	0.2917	0.5047	0.5929	0.3279
α_noDAE : α_DAEmut	0.8126	0.2524	0.6495	0.7574	0.3052	0.5637

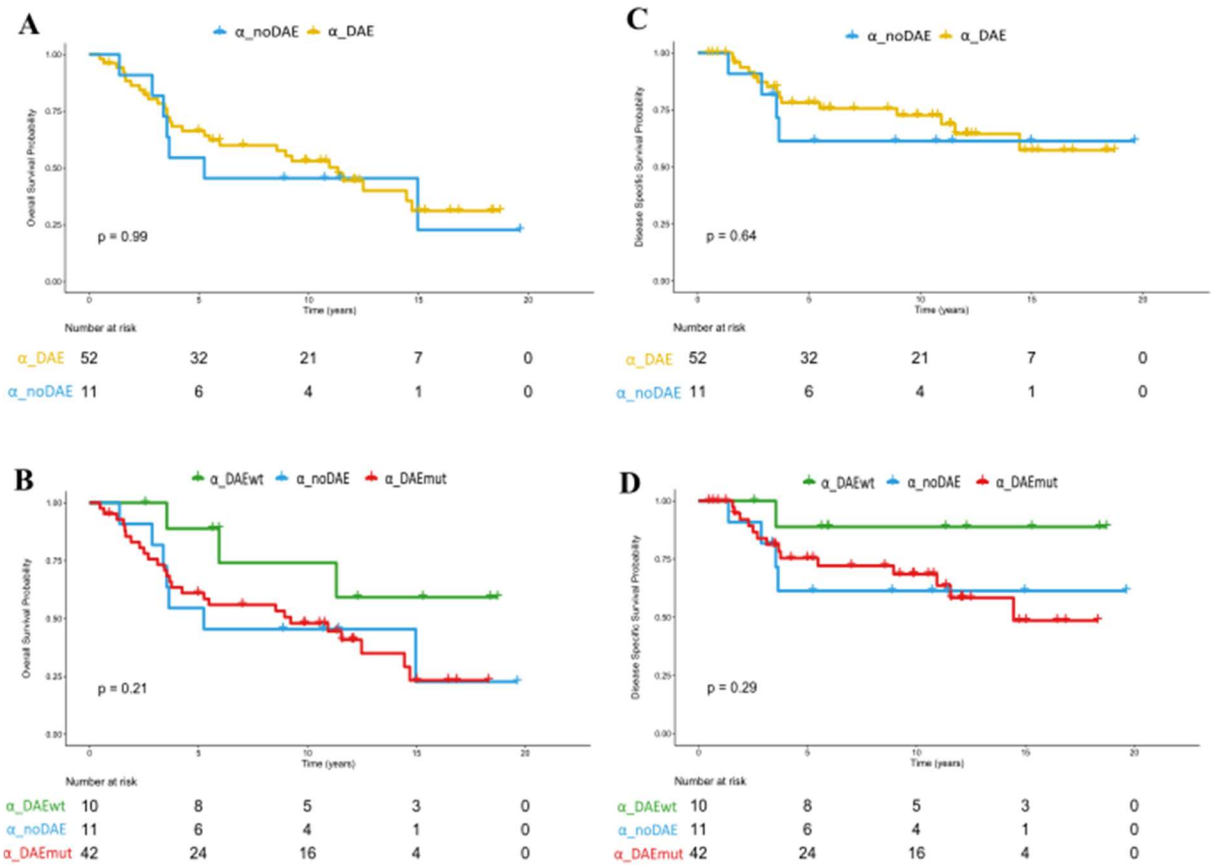


Figure T.8: Kaplan-Meier analysis of overall survival and disease specific survival of TP53's α_DAE groups in the METABRIC data set. (A) Overall Survival of METABRIC set divided in α_DAE and α_noDAE groups. (B) Overall Survival with α_DAE group further divided in tumours that express more wild-type allele (α_DAEwt) and tumours that express more mutated allele (α_DAEmut). (C) Disease Specific Survival of METABRIC set divided in α_DAE and α_noDAE groups. (D) Disease Specific Survival with α_DAE group further divided in tumours that express more wild-type allele (α_DAEwt) and tumours that express more mutated allele (α_DAEmut).

Table T.4: P-values of pairwise comparison overall and disease specific survivals of *TP53's* α_DAE groups in METABRIC set (p-values calculated by log-rank test if curves do not cross and by Two-Stage procedure if curves cross).

	OS	DSS
α_DAEwt : α_noDAE	0.1233	0.1335
α_DAEwt : α_DAEmut	0.089	0.4292
α_noDAE : α_DAEmut	0.9616	0.6435

Annex U

Table U.1: Summary Statistic of *TP53's* MADE Survival Analysis in TCGA set (Median, LCL and UCL in days; LCL: lower confidence level; UCL: uper confidence level; N: number of people of each group).

		Overall Survival					Disease Specific Survival				
		N	**Events**	**Median**	**0.95LCL**	**0.95UCL**	**N**	**Events**	**Median**	**0.95LCL**	**0.95UCL**
All	MADE	120	26	2192	1365	NA	119	18	NA	NA	NA
	mut=wt	7	1	1556	NA	NA	6	0	NA	NA	NA
ER positive	MADE	53	12	2192	1365	NA	53	6	NA	NA	NA
	mut=wt	4	0	NA	NA	NA	4	0	NA	NA	NA
ER negative	MADE	62	11	NA	1688	NA	61	10	NA	NA	NA
	mut=wt	3	1	1556	NA	NA	2	0	NA	NA	NA
PR positive	MADE	47	7	2192	959	NA	50	5	NA	NA	NA
	mut=wt	7	3	2273	1481	NA	4	0	NA	NA	NA
PR negative	MADE	65	13	NA	1365	NA	64	11	NA	NA	NA
	mut=wt	3	1	1556	NA	NA	2	0	NA	NA	NA
HER2 positive	MADE	29	2	NA	NA	NA	29	1	NA	NA	NA
	mut=wt	0	—	—	—	—	0	—	—	—	—
HER2 negative	MADE	73	14	NA	1365	NA	72	9	NA	NA	NA
	mut=wt	7	1	1556	NA	NA	6	0	NA	NA	NA

Table U.2: Summary Statistic of *TP53's* MADE groups Survival Analysis in TCGA set (Median, LCL and UCL in days; LCL: lower confidence level; UCL: uper confidence level; N: number of people of each group).

		Overall Survival					Disease Specific Survival				
		N	**Events**	**Median**	**0.95LCL**	**0.95UCL**	**N**	**Events**	**Median**	**0.95LCL**	**0.95UCL**
All	MADE_wt	16	3	NA	754	NA	16	3	NA	774	NA
	mut=wt	7	1	1556	NA	NA	6	0	NA	NA	NA
	MADE_mut	104	23	2192	1481	NA	103	15	NA	NA	NA
ER positive	MADE_wt	6	0	NA	NA	NA	6	0	NA	NA	NA
	mut=wt	4	0	NA	NA	NA	4	0	NA	NA	NA
	MADE_mut	47	12	2192	1365	NA	47	6	NA	NA	NA
ER negative	MADE_wt	10	3	785	754	NA	10	3	NA	774	NA
	mut=wt	3	1	1556	NA	NA	2	0	NA	NA	NA
	MADE_mut	52	8	NA	1688	NA	51	7	NA	NA	NA
PR positive	MADE_wt	6	0	NA	NA	NA	6	0	NA	NA	NA
	mut=wt	4	0	NA	NA	NA	4	0	NA	NA	NA
	MADE_mut	44	10	2192	1481	NA	44	5	NA	NA	NA
PR negative	MADE_wt	9	2	NA	754	NA	9	2	NA	774	NA
	mut=wt	3	1	1556	NA	NA	2	0	NA	NA	NA
	MADE_mut	56	11	NA	1365	NA	55	9	NA	NA	NA
HER2 positive	MADE_wt	7	0	NA	NA	NA	7	0	NA	NA	NA
	mut=wt	0	—	—	—	—	0	—	—	—	—
	MADE_mut	22	2	NA	959	NA	22	1	NA	NA	NA
HER2 negative	MADE_wt	8	3	785	754	NA	8	3	774	744	NA
	mut=wt	7	1	1556	NA	NA	6	0	NA	NA	NA
	MADE_mut	65	11	NA	1365	NA	64	6	NA	NA	NA

Table U.3: Summary Statistic of *TP53's* α_DAE groups Survival Analysis in TCGA set (Median, LCL and UCL in days; LCL: lower confidence level; UCL: uper confidence level; N: number of people of each group).

		Overall Survival					Disease Specific Survival				
		N	**Events**	**Median**	**0.95LCL**	**0.95UCL**	**N**	**Events**	**Median**	**0.95LCL**	**0.95UCL**
All	α_DAE	116	23	2192	1174	NA	115	17	NA	NA	NA
	α_noDAE	11	4	2273	1481	NA	10	1	NA	NA	NA
ER positive	α_DAE	50	9	2192	1034	NA	50	5	NA	2163	NA
	α_noDAE	7	3	2273	1481	NA	7	1	NA	NA	NA
ER negative	α_DAE	61	11	NA	967	NA	60	10	NA	NA	NA
	α_noDAE	4	1	1556	NA	NA	3	0	NA	NA	NA
PR positive	α_DAE	47	7	2192	959	NA	47	4	2163	NA	NA
	α_noDAE	7	3	2273	1481	NA	7	1	NA	NA	NA
PR negative	α_DAE	64	13	NA	1365	NA	63	11	NA	NA	NA
	α_noDAE	4	1	1556	NA	NA	3	0	NA	NA	NA
HER2 positive	α_DAE	28	2	NA	959	NA	28	1	NA	NA	NA
	α_noDAE	1	0	NA	NA	NA	1	0	NA	NA	NA
HER2 negative	α_DAE	72	12	NA	967	NA	71	9	NA	NA	NA
	α_noDAE	8	3	1556	1481	NA	7	0	NA	NA	NA

Table U.4: Summary Statistic of *TP53's* α_DAE grops Survival Analysis in TCGA set (Median, LCL and UCL in days; LCL: lower confidence level; UCL: uper confidence level; N: number of people of each group).

		Overall Survival					Disease Specific Survival				
		N	**Events**	**Median**	**0.95LCL**	**0.95UCL**	**N**	**Events**	**Median**	**0.95LCL**	**0.95UCL**
All	α_DAEwt	21	4	1034	785	NA	21	4	NA	1020	NA
	α_noDAE	11	4	2273	1481	NA	10	1	NA	NA	NA
	α_DAEmut	95	19	2192	1365	NA	94	13	NA	NA	NA
ER positive	α_DAEwt	9	1	1034	1034	NA	9	1	NA	1020	NA
	α_noDAE	7	3	2273	1481	NA	7	1	NA	NA	NA
	α_DAEmut	41	8	2192	1362	NA	41	4	NA	NA	NA
ER negative	α_DAEwt	12	3	785	754	NA	12	3	NA	774	NA
	α_noDAE	4	1	1556	NA	NA	3	0	NA	NA	NA
	α_DAEmut	49	8	NA	1688	NA	48	7	NA	NA	NA
PR positive	α_DAEwt	9	0	NA	NA	NA	9	0	NA	NA	NA
	α_noDAE	7	3	2273	1481	NA	7	1	NA	NA	NA
	α_DAEmut	38	7	2192	959	NA	38	4	NA	2163	NA
PR negative	α_DAEwt	11	3	1034	754	NA	11	3	NA	1020	NA
	α_noDAE	4	1	1556	NA	NA	3	0	NA	NA	NA
	α_DAEmut	53	10	NA	1365	NA	52	8	NA	NA	NA
HER2 positive	α_DAEwt	7	0	NA	NA	NA	7	0	NA	NA	NA
	α_noDAE	1	0	NA	NA	NA	1	0	NA	NA	NA
	α_DAEmut	21	2	NA	959	NA	21	1	NA	NA	NA
HER2 negative	α_DAEwt	12	3	785	754	NA	12	3	NA	774	NA
	α_noDAE	8	3	1556	1481	NA	7	0	NA	NA	NA
	α_DAEmut	60	9	NA	1365	NA	59	6	NA	NA	NA

Annex V – *TP53's* Kaplan-Meier Survival Analysis for TCGA set

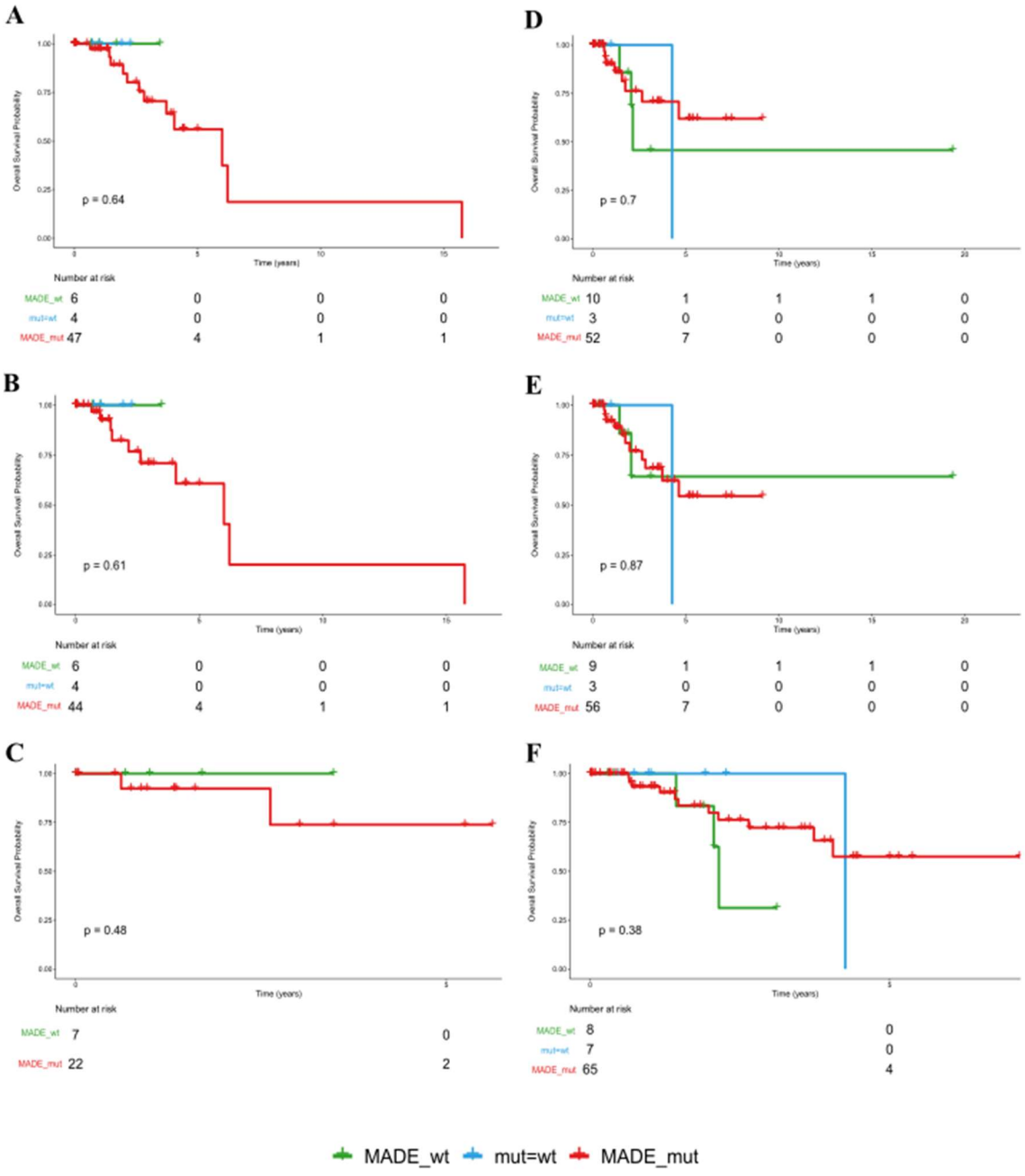


Figure V.1: Kaplan-Meier analysis of overall survival of *TP53's* MADE groups in the TCGA set according to ER, PR and HER2 statuses. Overall survival in (A) ER positive group, (B) PR positive group, (C) HER2 positive group, (D) ER negative group, (E) PR negative group and (F) HER2 negative group.

Table V.1: P-values of pairwise comparison of overall survival of *TP53's* MADE in TCGA set (p-values calculated by log-rank test if curves do not cross and by Two-Stage procedure if curves cross).

	ER positive	ER negative	PR positive	PR negative	HER2 positive	HER2 negative
MADE_wt : mut=wt	—	0.528	—	0.7092		0.0254
MADE_wt : MADE_mut	0.1355	0.3468	0.1965	0.6547	0.48	0.1635
mut=wt : MADE_mut	0.0418	0.4311	0.4568	0.0946		0.1957

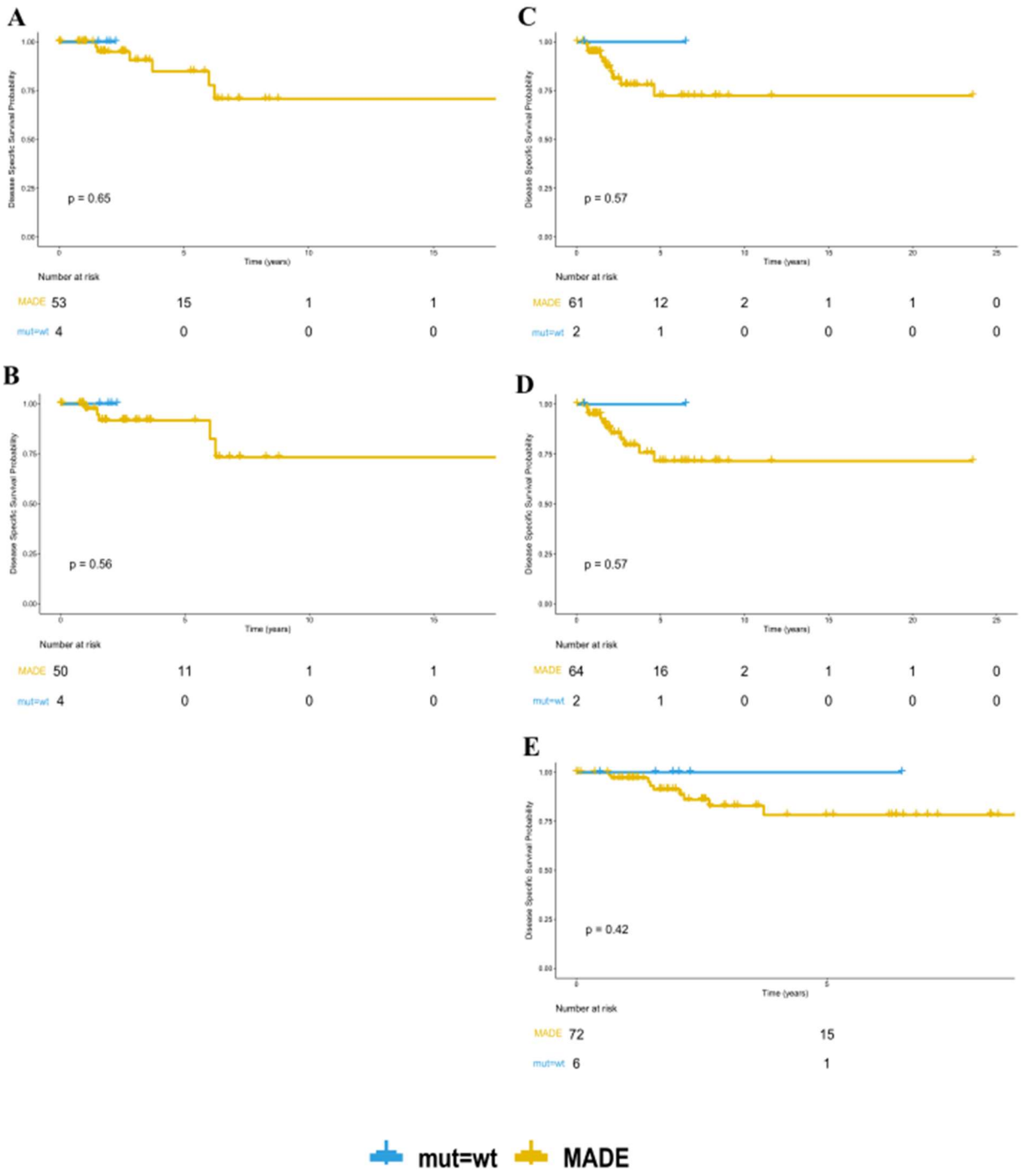


Figure V.2: Kaplan-Meier analysis of disease specific survival of *TP53's* MADE in the TCGA set according to ER, PR and HER2 statuses. Disease specific survival in (A) ER positive group, (B) PR positive group, (C) ER negative group, (D) PR negative group and (E) HER2 negative group.

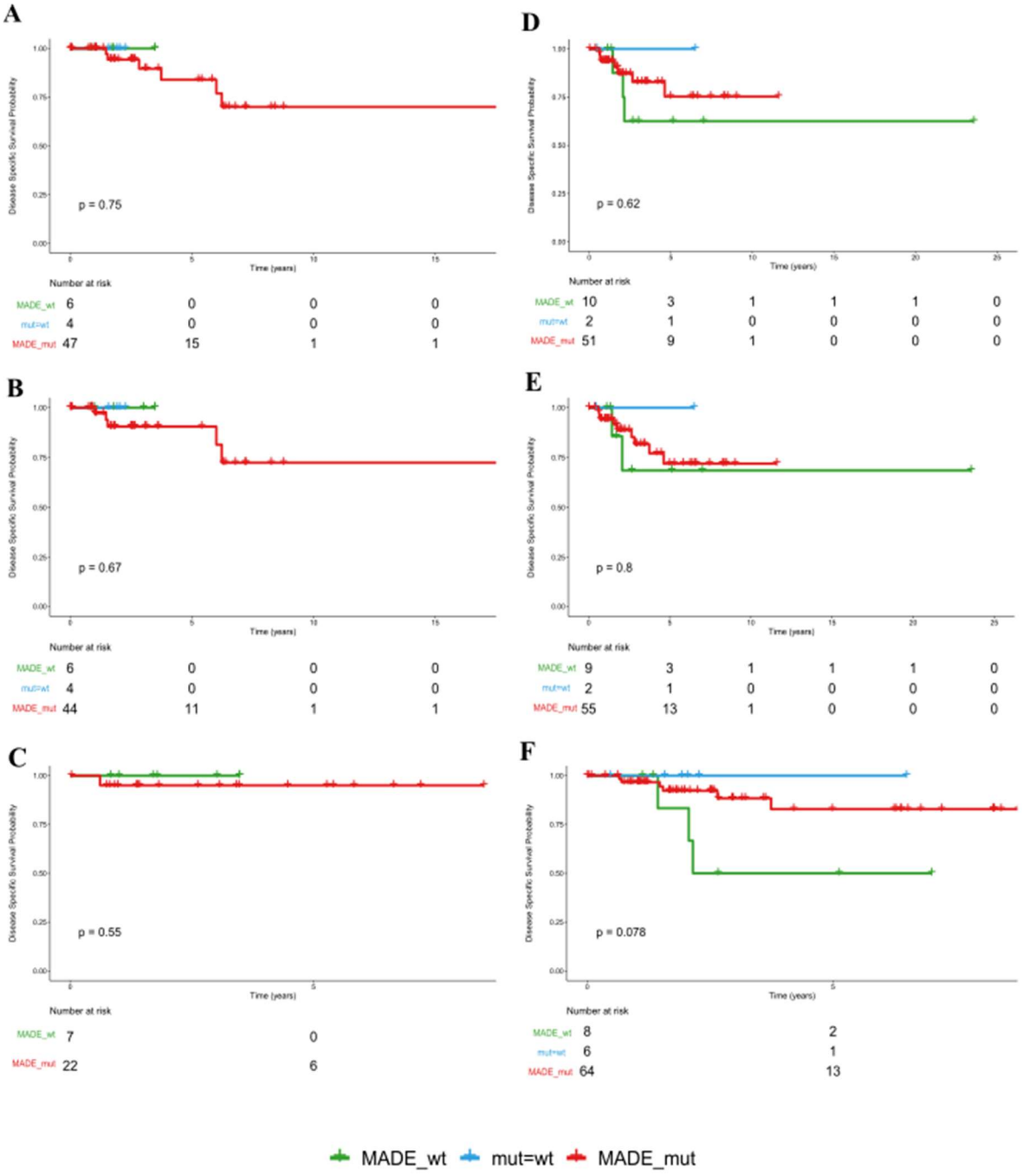


Figure V.3: Kaplan-Meier analysis of disease specific survival of *TP53's* MADE groups in the TCGA set according to ER, PR and HER2 statuses. Disease specific survival in (A) ER positive group, (B) PR positive group, (C) HER2 positive group, (D) ER negative group, (E) PR negative group and (F) HER2 negative group.

Table V.2: P-values of pairwise comparison of disease specific survival of *TP53's* MADE in TCGA set (p-values calculated by log-rank test if curves do not cross and by Two-Stage procedure if curves cross).

	ER positive	ER negative	PR positive	PR negative	HER2 positive	HER2 negative
MADE_wt : mut=wt	—	0.5098	—	0.5582		0.11
MADE_wt : MADE_mut	0.564	0.5456	0.5171	0.8246	0.55	0.8924
mut=wt : MADE_mut	0.6301	0.5994	0.5313	0.7189		0.4774

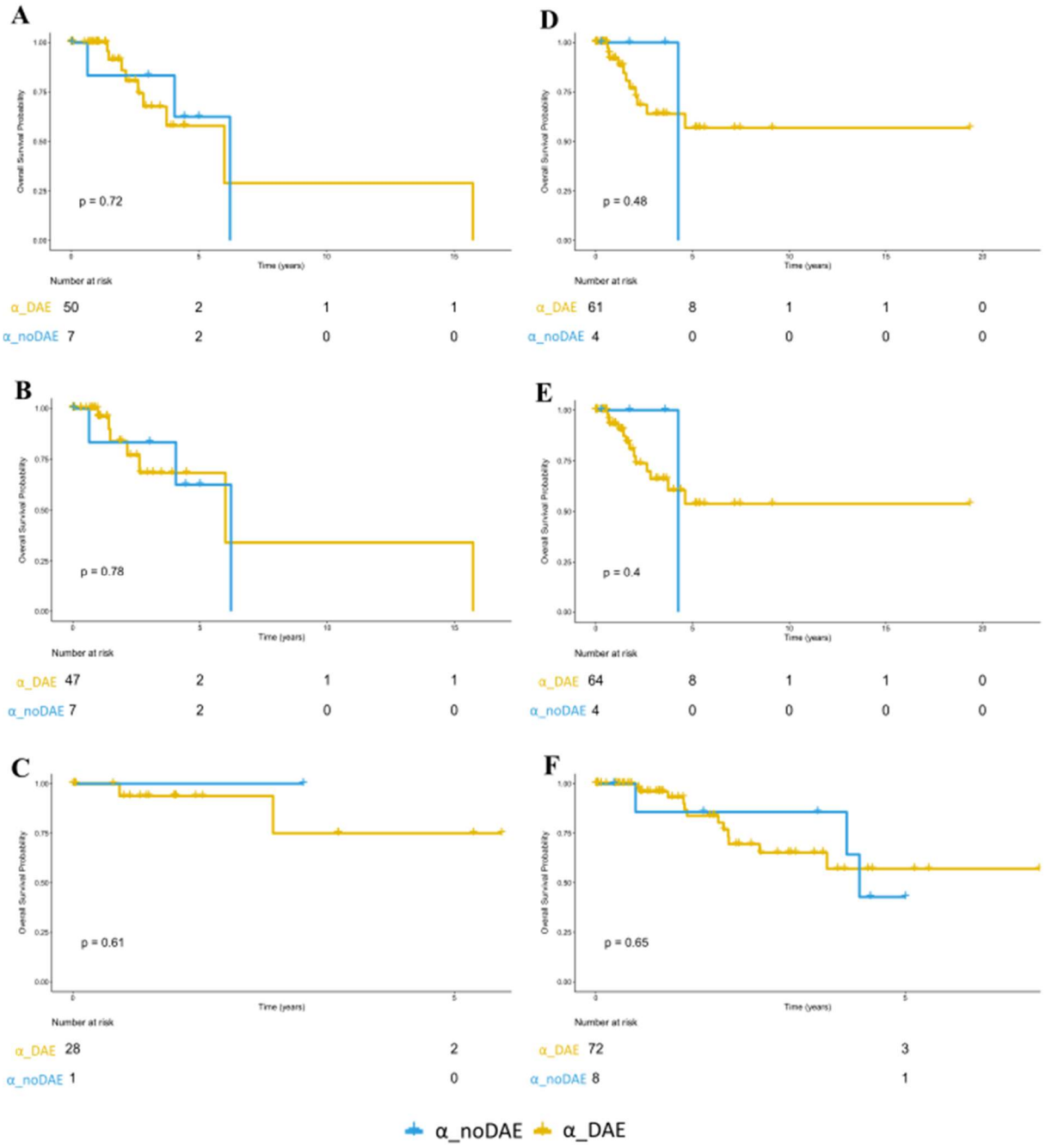


Figure V.4: Kaplan-Meier analysis of overall survival of *TP53's* α_DAE groups in the TCGA set according to ER, PR and HER2 statuses. Overall survival in (A) ER positive group, (B) PR positive group, (C) HER2 positive group, (D) ER negative group, (E) PR negative group and (F) HER2 negative group.

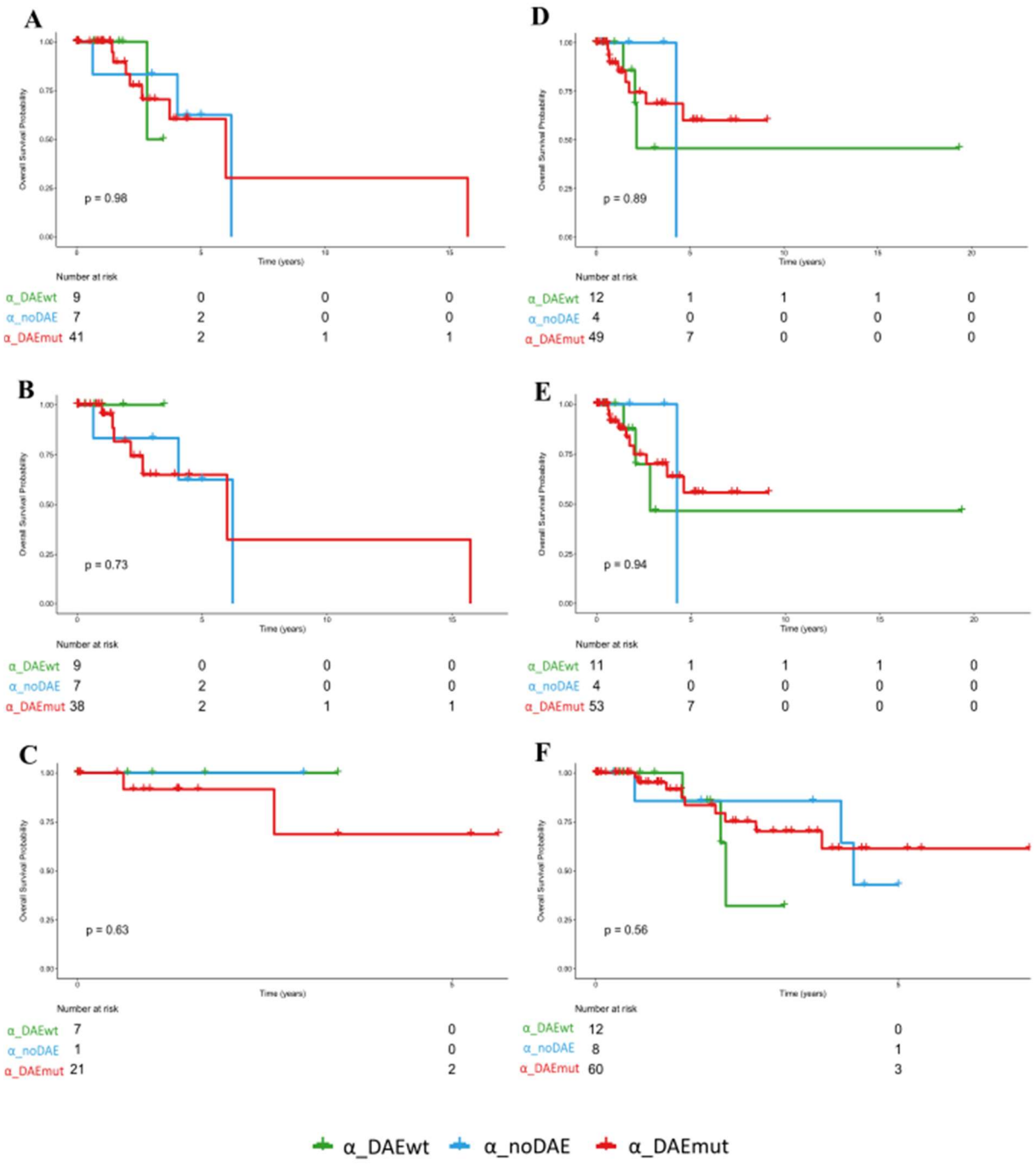


Figure V.5: Kaplan-Meier analysis of overall survival of *TP53's* α_DAE groups in the TCGA set according to ER, PR and HER2 statuses. Overall survival in (A) ER positive group, (B) PR positive group, (C) HER2 positive group, (D) ER negative group, (E) PR negative group and (F) HER2 negative group.

Table V.3: P-values of pairwise comparison of overall survival of *TP53's* α_DAE groups in TCGA set (p-values calculated by log-rank test if curves do not cross and by Two-Stage procedure if curves cross).

	ER positive	ER negative	PR positive	PR negative	HER2 positive	HER2 negative
α_DAEwt : α_noDAE	0.9853	0.4181	0.3613	0.4207	—	0.4207
α_DAEwt : α_DAEmut	0.3795	0.3349	0.4182	0.2867	0.4452	0.2357
α_noDAE : α_DAEmut	0.79	0.6793	0.7787	0.5137	0.5645	0.7148

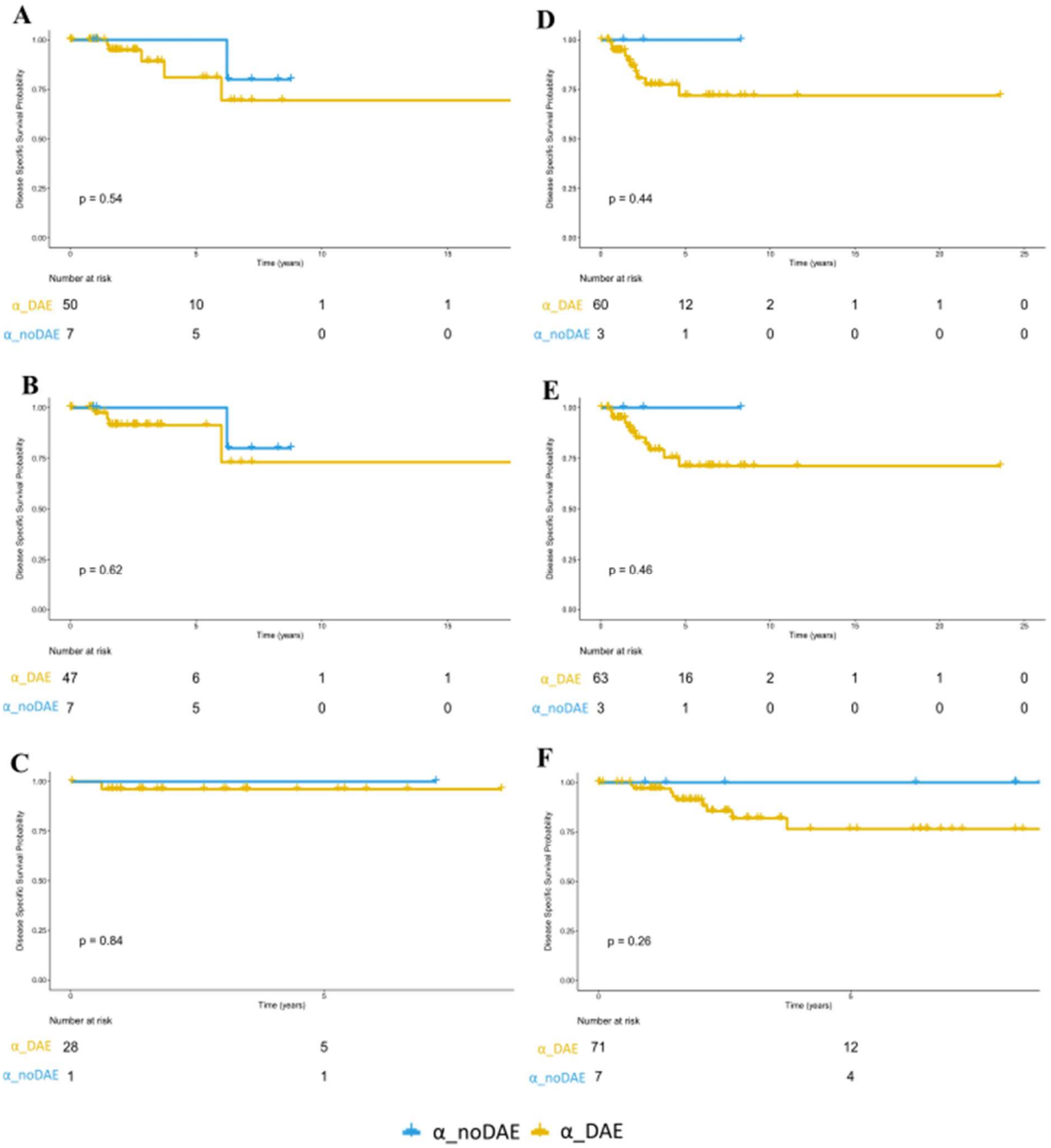


Figure V.6: Kaplan-Meier analysis of disease specific survival of *TP53's* α_DAE groups in the TCGA set according to ER, PR and HER2 statuses. Disease specific survival in (A) ER positive group, (B) PR positive group, (C) HER2 positive group, (D) ER negative group, (E) PR negative group and (F) HER2 negative group.

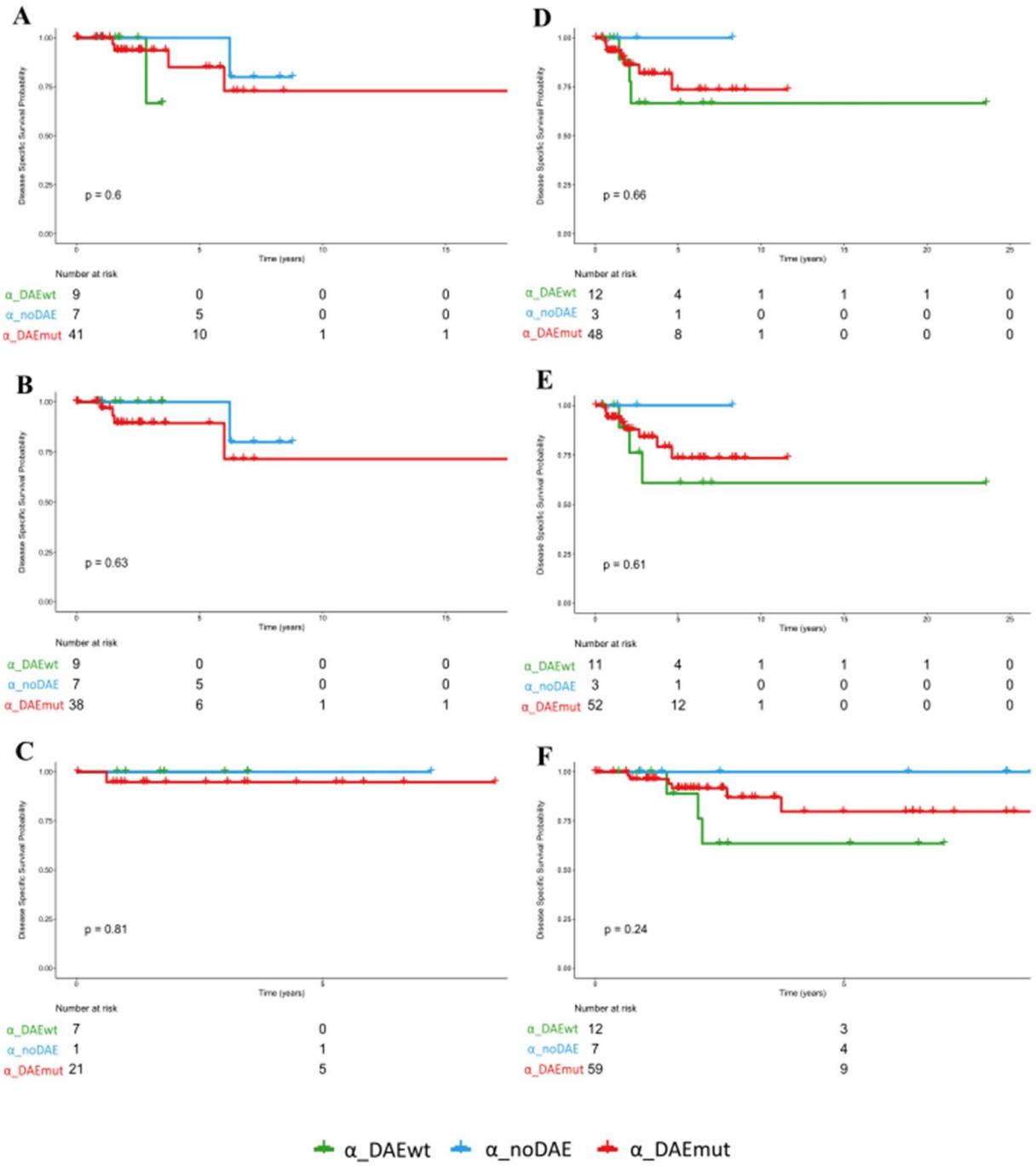


Figure V.7: Kaplan-Meier analysis of disease specific survival of *TP53's* α_DAE groups in the TCGA set according to ER, PR and HER2 statuses. Disease specific survival in (A) ER positive group, (B) PR positive group, (C) HER2 positive group, (D) ER negative group, (E) PR negative group and (F) HER2 negative group.

Table V.4: P-values of pairwise comparison of disease specific survival of *TP53's* α_DAE groups in TCGA set (p-values calculated by log-rank test if curves do not cross and by Two-Stage procedure if curves cross).

	ER positive	ER negative	PR positive	PR negative	HER2 positive	HER2 negative
α_DAEwt : α_noDAE	0.6982	0.3841	—	0.4003	—	0.1478
α_DAEwt : α_DAEmut	0.7992	0.5597	0.4063	0.5113	0.5438	0.8733
α_noDAE : α_DAEmut	0.5293	0.4743	0.569	0.4833	0.8185	0.3086

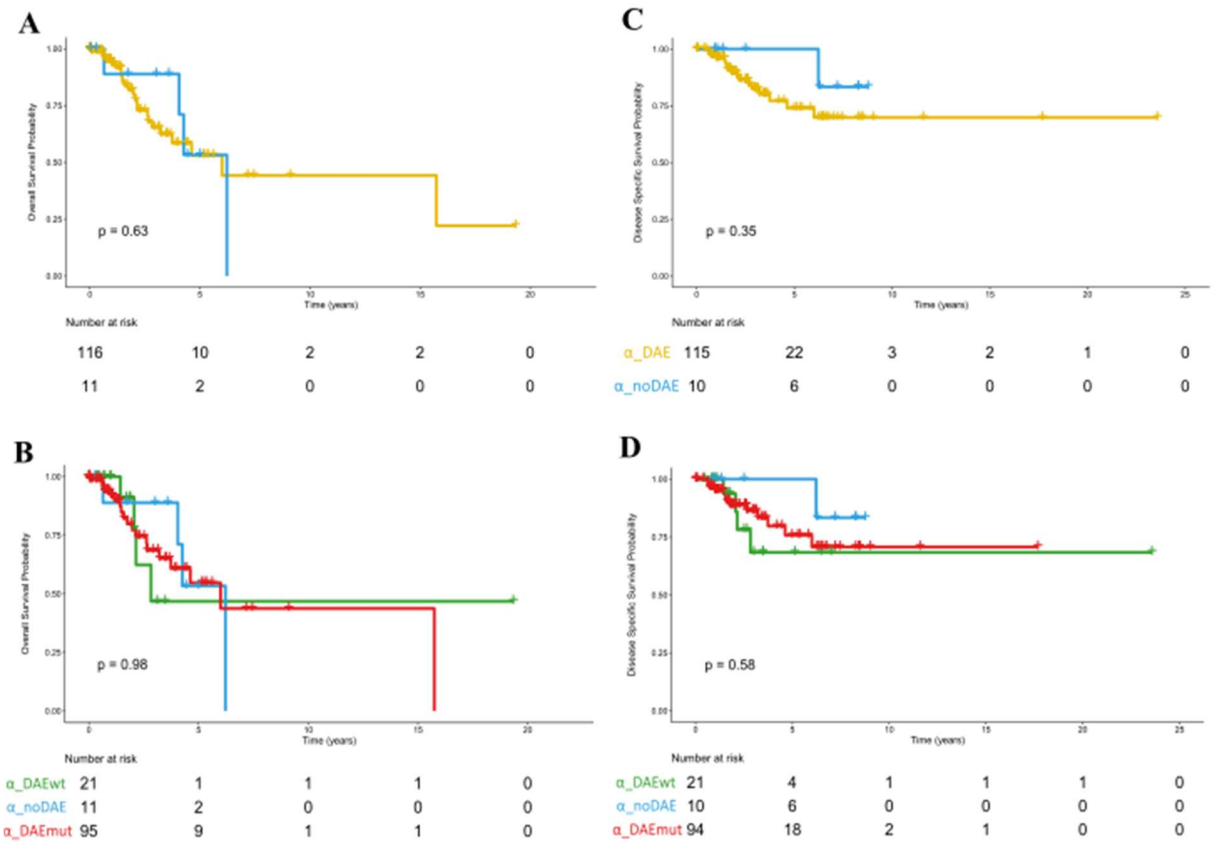


Figure V.8: Kaplan-Meier analysis of overall survival and disease specific survival of TP53's α_DAE groups in the TCGA data set. (A) Overall Survival of TCGA set divided in α_DAE and α_noDAE groups. (B) Overall Survival with α_DAE group further divided in tumours that express more wild-type allele (α_DAEwt) and tumours that express more mutated allele (α_DAEmut). (C) Disease Specific Survival of TCGA set divided in α_DAE and α_no_DAE groups. (D) Disease Specific Survival with α_DAE group further divided in tumours that express more wild-type allele (α_DAEwt) and tumours that express more mutated allele (α_DAEmut).

Table V.5: P-values of pairwise comparison of overall and disease specific survivals of *TP53's* α_DAE group in TCGA set (p-values calculated by log-rank test if curves do not cross and by Two-Stage procedure if curves cross).

	OS	DSS
α_DAEwt : α_noDAE	0.9208	0.2805
α_DAEwt : α_DAEmut	0.4059	0.438
α_noDAE : α_DAEmut	0.5836	0.9193