Prediction of High-Risk Cancer Patients using Clinical Factors and Expression Profile of Apoptosis Regulators

By

Chakit Arora

Under the Supervision of Prof. Gajendra P.S. Raghava

Department of Computational Biology

Indraprastha Institute of Information Technology Delhi

New Delhi - 110020

August, 2021

Prediction of High-Risk Cancer Patients using Clinical Factors and Expression Profile of Apoptosis Regulators

By

Chakit Arora

A Thesis

Submitted in Partial Fulfillment of the Requirements for the Degree Of

Doctor of Philosophy

Department of Computational Biology

Indraprastha Institute of Information Technology Delhi

New Delhi - 110020

August, 2021

Certificate

This is to certify that the thesis titled "**Prediction of High-Risk Cancer Patients using Clinical Factors and Expression Profile of Apoptosis Regulators**" being submitted by **Mr. Chakit Arora** to the Indraprastha Institute of Information Technology Delhi, for the award of the degree of Doctor of Philosophy, is an original research work carried out by him under my supervision. In my opinion, the thesis has reached the standards fulfilling the requirements of the regulations relating to the degree.

The results contained in this thesis have not been submitted in part or full to any other university or institute for the award of any degree/diploma.

August, 2021
Month, Year

Prof. Gajendra P.S. Raghava
Supervisor Name

Indraprastha Institute of Information Technology
New Delhi-110020, India

Declaration

I, **Chakit Arora**, hereby declare that I have not plagiarised any content in the thesis from an earlier source and the writing is my own. I also declare that no scientific misconduct is involved in the work. Each result included in this thesis has been reproduced by me at least twice using the dataset, scripts, and codes that were publicly submitted on the GitHub page ('raghavagps/Chakit_Thesis') or are accessible publicly at other platforms.

Chakit Arora

August, 2021

Acknowledgement

"Remember to look up at the stars and not down at your feet. Try to make sense of what you see and wonder about what makes the universe exist. Be curious. And however difficult life may seem, there is always something you can do and succeed at. It matters that you don't just give up." – Prof. Stephen Hawking

Transitioning from a different discipline was a very challenging task for me. There has been a significant contribution of a number of people who made this challenge easy for me. Hereby, I'd like to express my gratitude to the people who made it possible and due to whom I never gave up, however hectic the journey seemed.

First and foremost, I would like to acknowledge my indebtedness and give my warmest gratitude to Professor Gajendra Pal Singh Raghava, my advisor, for making this doctoral thesis possible. His friendly encouragement and professional suggestions were indispensable throughout all phases of the work. His unwavering enthusiasm for science kept me engrossed in my research, and his personal generosity helped me thrive through my doctoral journey at IIITD. The immense support which he provided whenever I was devoid of any resorts is really admirable. There are many things he possesses that I'd like to incorporate in my psyche especially his perspective about life and how to live it. The discussions we had will always be in my memories and would inspire me to live life more practically and cheerfully. One thing that I'll always recall is his childhood story where he used to pick locks much like what the Nobel laureate Richard Feynman used to do during his childhood. I believe this is his greatest strength, the art of changing "frames of reference" and then solving a problem from different angles - "Think how the thief would think!" - as he often says. Lastly, I'd like to appreciate the friendly environment he imposes and motivates, wherein both students and PIs are encouraged to collaborate and learn from each other.

I sincerely thank Drs. Subhadip Raychaudhuri, Ganesh Bagler and Tavpritesh Sethi for the mentoring they provided and constructive support throughout this work. I am highly grateful to the current and ex-faculty of Department of Computational Biology for imparting their knowledge in the form of courses they taught as well as in the informal discussions, especially Dr. K Sriram, Dr. Ganesh Bagler and Dr. Arnab Bhattacharjee. I would also like to thank the instructors of the courses to whom I was appointed as a teaching assistant, especially Dr. Raj Ayyar, whose cheerful persona, philosophical insights and free chocolates/coffees/pizzas was a delight to both my mental

and otherwise appetite. It was also highly refreshing to discuss ‘Quantum Mechanics’ with Dr. Sukanta Dutta for the brief time he was here. Ultimately, I would like to thank the administrative staff of IIITD especially Mrs. Priti Patel, Ms. Shipra Jain and Ms. Sheetu Ahuja for consistently being there to manage our queries and handling the academic affairs so diligently. I am also grateful to IIITD for lending financial support along-with the first-class infrastructure and facilities.

My appreciation extends to the friends I made at IIITD and laboratory colleagues. This journey wouldn’t have been complete without a lot of people, although, it is impossible to mention everyone here. Therefore, I’d like to exclude a lot of names (ex-Btech and ex-Mtech students of IIITD), who were possibly the most important and at the same time are unlikely to ever read this thesis. In particular, I’d like to thank my friends and seniors Dr. Pawan Kumar Raghav, Dr. Akshara, Dr. Sherry, Dr. Piyush, Dr. Salman, Dr. Rajesh and Dr. Harpreet for their support and assistance whenever required. I am also grateful to my friends and colleagues Dilraj Kaur, Dr. Anjali Lathwal, Vinod, Anjali Dhall, Sumeet, Neelam, Akanksha, Ritu, and Dr. Devi. From the friendly neighbours, I’d like to express thanks to Omkar, Neetesh, Raghav, Smriti, Sarita, Krishan, Shreya, Chitrita, Priyadarshini, Aditya and Dr. Shiju. The numerous experiences I share with each of them will always be memorable.

It’s difficult to adequately describe the kind of crucial life support those closest to me have provided. I hope they would extrapolate from the few words I have managed to put down. A special note of thanks to my wife Dilraj Kaur for her love and understanding. She has been a great friend, a colleague and a driving force throughout my journey here. I am extremely grateful to my parents Mr. Bharat Bhushan Arora and Mrs. Anju Arora, for their undaunted trust in me and my decisions. Lastly, I’d like to acknowledge my sister Himani Arora and my brother in law Alok Nahata for always being patient listeners and encouraging me in every step of my life.

Chakit Arora

Abstract

With about 19 million occurrences and 10 million mortalities in 2020, “Cancer” is the second leading source of mortality worldwide (WHO GLOBOCAN). The top three continents burdened with cancer deaths are Asia (58.3 percent), Europe (19.6 percent) and Latin America (7.2 percent). As of today, patient management in cancer care involves three broad steps: (a) screening and diagnosis, (b) risk assessment and prognosis, and (c) therapy. Since therapeutic intervention follows the risk assessment step, it is known to be the most critical phase in the cancer care and treatment. Risk estimation is done by means of multiple staging schemes for most cancers. The 'TNM system' for which the staging directives are issued by the “American Joint Committee on Cancer” (AJCC) and the “Union for International Cancer Control” (UICC), is the most extensively used system. The overall stage in the TNM system is determined when a letter (often with a number) is allocated to the cancer to describe the stages of T: tumour, N: node and M: metastasis , in which T specifies the size and location of the initial tumour, N indicates cancer spread to the adjacent lymph nodes, and M shows the cancer spread to distant body parts. The traditional TNM staging only involved anatomical considerations, but the modern staging system is continuously revised to provide details on other characteristics such as cancer biomarkers that include the profile/status of certain molecules that are altered in cancer tissues and clinical characteristics such as the location of tumour or age. These insights are integrated into the staging processes for various kinds of cancer, which makes it more reliable and useful to both doctors and patients. For example the recent inclusion of HER2 status was a result of a new Neo-Bioscore staging system, thereby allowing more precise prognostic stratification of all breast cancer subtypes . The addition of ‘Age’ in Thyroid cancer staging has also been reported to improve risk assessment.

The heterogeneity associated with cancer is a major hurdle in the formulation of “cancer biomarkers”, as each cancer is comprised of multiple phenotypes and frequently responds differently to the same therapeutic intervention. This heterogeneity exists because of the aberrant behaviour of cancer cells, not just in different types of cancer, but even in same cancer type. In order to resolve this and persuade a “personalized medicine” approach, modern oncologists are actively seeking to develop a thorough understanding of the molecular processes that drive cancer. Biomarker development using genomic and proteomic data is now considered to be a superior means of carefully approaching the problem of cancer heterogeneity. This is largely achieved by a detailed study of data obtained from subcellular processes that drive oncogenesis. In this study, we focused on a prominent cellular pathway, Apoptosis, which has a strong and proven background in the growth and development of cancer. In the framework of genomic data, for the

particular case of thyroid cancer, we demonstrate that certain genes belonging to the apoptotic pathway are associated with patient survival. The elevation and suppression of mRNA levels of these genes may be responsible for an aggressive or a mild phenotype of thyroid cancer thereby affecting patient outcome. The proposed signature in a further analysis was shown to perform better than AJCC staging, for risk stratification purposes,. The identified genes also exhibit a differential expression between normal and cancerous tissue, suggesting their ability to distinguish between individuals with and without cancer. Further, it was shown that the application of a similar approach to a pan-cancer analysis revealed universal gene signatures that have prognostic significance across various cancer types. This is in contrast with the conventional cancer-specific biomarker development process. The study centred at identification of prognostic biomarker and devised a 11 gene panel that is applicable across 27 cancer types. Although, the panel's efficiency is seen to differ among cancer types, a substantial stratification is achieved in all cases. In addition to this, the study provides a new cross cancer biomarker development approach and sheds light on a new gene signature that can be used in patients with brain or kidney cancer. In the area of cancer treatment and rehabilitation, the practical realisation of such versatile biomarkers poses enormous benefits.

Gene expression profiling is a very accurate strategy for the understanding of cancer and its prognosis, but, in the context of signalling networks, the activity of these genes depends on their translation into functional proteins. Because fundamental protein families controlling the apoptotic pathway together with their roles are commonly known, an in-depth study of the proteomic profiles of different tissues retrieved from cancer diagnosed individuals is anticipated to enhance our comprehension of tumour pathogenesis, prognosis, and recognition of therapeutic targets. To this end, the analysis included a proteomic dataset with the expression profile of Bcl2 family proteins in the scope of colo-rectal cancer. Information from previous apoptotic pathway studies has been used to establish a predictive biomarker for the estimation of response to treatment in colorectal cancer patients. This research illuminated the synergistic function of proteins in conferring therapeutic "resistance" to colorectal cancer and the critical role of apoptosis. The prognostic power of the biomarker was compared to different clinical features and methods. The method was released into public domain by the means of a web-server, thereby enforcing its practical utility to both researchers and clinicians.

However, a major problem with biomarkers focused on "omics" is that inclusion of these biomarkers makes staging processes more complicated, rendering them difficult for people to

understand. Thus, considering their outstanding success in cohort trials, most biomarkers have not yet been applied to the staging schemes. Therefore, our current research also examines the importance of numerous 'clinical factors/features' that collectively include pathological features, demographic characteristics, lifestyle-related features, anatomical characteristics, blood protein status (such as ER) in evaluating cancer patients' survival outcomes. Apart from the comprehensive assessment of clinical factors and their integration to the gene/protein signatures proposed above, we explicitly studied the case of "Melanoma" and looked at the prognostic power of genomic information pertaining to many cancer-associated pathways as well as clinical factors. We demonstrate that a prognostic model that incorportates only clinical factors is superior to the model focused on gene expression. This research also illustrates the significance of clinical factors for risk assessment. It shows how the schematic incorporation of existing clinical features into the staging process can be more successful. It also indicates that while omics-based biomarkers could be desirable due to their inherent biological correlation, clinical factors should not be undermined. On the basis of this pretext, the study is further expanded to the pan-cancer framework of designing risk prediction models by using only clinical factors. The clinical factors concerned include a wide variety of characteristics, ranging from inherent or heritable factors, different extrinsic risk factors, physiological features and surgical or therapeutic procedures used. The study established risk prediction models that are easy to apply and understand. Models were also evalauted against staging systems in various cancer cohorts.

Overall, the study discussed in this thesis suggests some novel prognostic biomarkers and approaches for improved risk management in cancer patients. On the one side, the pipeline used in the analysis exploited a key cellular mechanism by using recent "omics"-based information. On the other hand, different clinical factors were examined both independently and in conjunction with proposed biomarker genes/proteins in regard to patient survival. The study discussed here can be useful for the development of better therapeutic modalities and thereby aid in the advancement of cancer research.

List of Publications

Thesis Publications

- Lathwal A*, **Arora C***, Raghava GPS. Prediction of risk scores for colorectal cancer patients from the concentration of proteins involved in mitochondrial apoptotic pathway. PLoS One. 2019. *14*(9), p.e0217527. ***joint first author***

- **Arora C**, Kaur D, Lathwal A, Raghava GPS. 2020. Risk prediction in cutaneous melanoma patients from their clinico-pathological features: superiority of clinical data over gene expression data. *Heliyon* **6**. https://doi.org/10.1016/j.heliyon.2020.e04811

- **Arora C**, Kaur D, Raghava GPS. Prognostic Biomarkers for Predicting Papillary Thyroid Carcinoma Patients at High Risk Using Nine Genes of Apoptotic Pathway. (***under review, PloS One***)

- **Arora C**, Kaur D, Raghava GPS. Universal Prognostic Biomarkers for Predicting Survival Risk of Cancer Patients from Expression Profile of Apoptotic Pathway Genes. (***under review, Wiley Proteomics***)

Other Publications

- Kaur, D., **Arora, C**. and Raghava, G.P.S., 2021. Prognostic Biomarker-Based Identification of Drugs for Managing the Treatment of Endometrial Cancer. *Molecular Diagnosis & Therapy*, pp.1-18.

- Lathwal, A., Kumar, R., Arora, C. and Raghava, G.P.S., 2020. Identification of prognostic biomarkers for major subtypes of non-small-cell lung cancer using genomic and clinical data. *Journal of Cancer Research and Clinical Oncology*, *146*(11), pp.2743-2752.

- Kaur, D.*, **Arora, C**.*, & Raghava, G. P. (2020). A hybrid model for predicting pattern recognition receptors using evolutionary information. *Frontiers in immunology*, *11*, 71. ***joint first author***

- Pande, A., Patiyal, S., Lathwal, A., **Arora, C**., Kaur, D., Dhall, A., Mishra, G., Kaur, H., Sharma, N., Jain, S. and Usmani, S.S., 2019. Computing wide range of protein/peptide features from their sequence and structure. *bioRxiv*, p.599126. (***in communication***)

- Dhall, A., Patiyal, S., Kaur, H., Bhalla, S., **Arora, C**., & Raghava, G. P. (2020). Computing Skin Cutaneous Melanoma Outcome from the HLA-alleles and Clinical Characteristics. *Frontiers in genetics*, *11*, 221.

- Sharma, N., Patiyal, S., Dhall, A., Pande, A., **Arora, C.** and Raghava, G.P., 2021. AlgPred 2.0: an improved method for predicting allergenic proteins and mapping of IgE epitopes. *Briefings in bioinformatics*, *22*(4), p.bbaa294.

- Patiyal, S., Kaur, D., Kaur, H., Sharma, N., Dhall, A., Sahai, S., Agrawal, P., Maryam, L., **Arora, C.** and Raghava, G.P., 2020. A web-based platform on Coronavirus Disease-19 to maintain predicted diagnostic, drug, and vaccine candidates. *Monoclonal antibodies in immunodiagnosis and immunotherapy*, *39*(6), pp.204-216.

Table of Contents

List of Abbreviations

ACC	Adrenocortical carcinoma
BLCA	Bladder Urothelial Carcinoma
LGG	Brain Lower Grade Glioma
BRCA	Breast invasive carcinoma
CESC	Cervical squamous cell carcinoma and endocervical adenocarcinoma
CHOL	Cholangiocarcinoma
LCML	Chronic Myelogenous Leukemia
COAD	Colon adenocarcinoma
ESCA	Esophageal carcinoma
GBM	Glioblastoma multiforme
HNSC	Head and Neck squamous cell carcinoma
KICH	Kidney Chromophobe
KIRC	Kidney renal clear cell carcinoma
KIRP	Kidney renal papillary cell carcinoma
LAML	Acute Myeloid Leukemia
LIHC	Liver hepatocellular carcinoma
LUAD	Lung adenocarcinoma
LUSC	Lung squamous cell carcinoma
DLBC	Lymphoid Neoplasm Diffuse Large B-cell Lymphoma
MESO	Mesothelioma
MISC	Miscellaneous
OV	Ovarian serous cystadenocarcinoma
PAAD	Pancreatic adenocarcinoma
PCPG	Pheochromocytoma and Paraganglioma
PRAD	Prostate adenocarcinoma
READ	Rectum adenocarcinoma
SARC	Sarcoma
SKCM	Skin Cutaneous Melanoma
STAD	Stomach adenocarcinoma
TGCT	Testicular Germ Cell Tumors
THYM	Thymoma
THCA	Thyroid carcinoma
UCS	Uterine Carcinosarcoma
UCEC	Uterine Corpus Endometrial Carcinoma

UVM	Uveal Melanoma
CRC	Colo-rectal cancer
TCGA	The cancer genome atlas
GEP	Gene expression profile
GO	Gene ontology
HR	Hazard ratio
GPM	Good prognostic marker
BPM	Bad prognostic marker
C	Concordance index
CI	Confidence interval
RS	Risk score
RG	Risk grade
Cox-PH	Cox proportional hazard
OS	Overall survival
AUROC	Area under receiver operating curve
PI	Prognostic index
RF	Random forest
SVR	Support vector regressor
KNN	k-nearest neighbours
DT	Decision Trees
MLR	Multiple linear regression
TP	True positive
FP	False positive
TN	True negative
FN	False negative
RSEM	RNA-Seq by Expectation-Maximization

List of Software/Databases

Name	Availability
sklearn	https://scikit-learn.org/stable/
caret	https://github.com/topepo/caret
survival	https://cran.r-project.org/web/packages/survival/index.html
survminer	https://cran.r-project.org/web/packages/survminer/index.html
survMisc	https://cran.r-project.org/web/packages/survMisc/index.html
TCGA-Assembler 2	https://github.com/compgenome365/TCGA-Assembler-2
Connectivity Map 2	https://clue.io/cmap
dendextend	https://cran.r-project.org/package=dendextend
survivalROC	https://cran.r-project.org/web/packages/survivalROC/index.html
Firehose	https://gdac.broadinstitute.org
randomForestSRC	https://cran.r-project.org/package=randomForestSRC
TCGA-GDC	https://portal.gdc.cancer.gov
AJCC Individualized Melanoma Prediction Tool	http://www.melanomaprognosis.net
StringDB	https://string-db.org
GEO	https://www.ncbi.nlm.nih.gov/geo/
GeneCards	https://www.genecards.org
CCGD	http://ccgd-starrlab.oit.umn.edu
GTEx	https://gtexportal.org/home/
GEPIA	http://gepia.cancer-pku.cn
SurvExpress	http://bioinformatica.mty.itesm.mx:8080/Biomatec/SurvivaX.jsp
HPA	https://www.proteinatlas.org
Programming environments	Python, R, HTML, PHP, CSS and Javascript

List of Figures

List of Tables

1

Introduction

1.1 Cancer

Cancer is the name of a group of more than a hundred diseases which share their fundamental origin in the malfunction of cellular machinery. Cellular organization of human tissues has made possible the creation of an unprecedented variety of anatomical designs. Most of this plasticity of architecture can be traced back to the fact that the human cells are endowed with great control and flexibility. Most forms of cells in the human body encapsulate a full human genome i.e. far more knowledge than any of these cells would ever need. Several cells maintain the capacity to expand and differentiate even after organism construction has been completed. The continued ability to proliferate and partake in tissue formation (morphogenesis) allows it to sustain adult tissues during the lifespan of the organism. Such maintenance may include the repair of defects and the reinforcement of cells that have suffered damage after long periods of action.

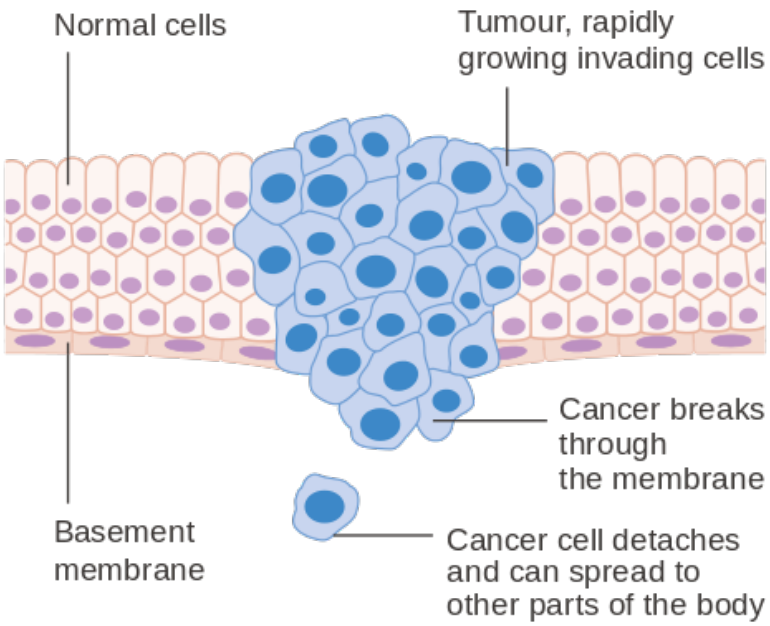


Figure 1.1 The growth and progression of Cancer. Cancer cells (both benign and malignant) multiply abnormally as compared to normal cells. Malignant tumour cells invade other body parts while benign are localized. (Image by Cancer Research UK, CC BY-SA 4.0, via Wikimedia Commons)

At the same time, this mobility and autonomy pose a significant problem, in that individual cells within the organism may obtain access to information that is typically inaccessible to them in their genes and perform functions that are inadequate for the preservation and operation of normal tissue. Furthermore, their genomic sequences are vulnerable to manipulation by various

mechanisms which modify the structure and, consequently, the information content of the genome. The resulting genetic mutations may redirect cells into the acquisition of new, often highly irregular phenotypes. Such modifications may be inconsistent with the standard roles of these cells in organismic function and biology. Modifications in cell growth and death programmes may be among these undesirable changes, and these may in turn contribute to the emergence of vast populations of cells that no longer follow the laws regulating natural tissue maintenance and repair. These rogue cells that form a 'tumour', also known as cancer cells, are the product of natural growth gone haywire. Cancer cells somehow learn to survive through exceptional precautions taken by the organism to deter their emergence. In creating the varied tissues that make organismic survival viable, normal cells are deliberately programmed to cooperate with each other. Cancer cells, on the other hand, have an agenda that is very different and more focused. They seem to be driven by only one consideration: creating more copies of themselves (**Figure 1.1**). Genetic alterations that lead to cancer appear to affect three major groups of genes—proto-oncogenes, tumour suppression genes, and DNA repair genes; which are primarily involved in controlling cell growth, repair and division processes. However, when these genes are altered in specific manner, they may become cancer-causing genes that cause cells to expand and survive when they are not permitted to do so.

Cancerous tumours are often categorized as benign or malignant tumours. Tumours that are malignant, can migrate to or infect surrounding tissues. (**Figure 1.1**) Benign tumors, on the other hand, do not grow into, or enter, surrounding tissues. A cancer that has expanded from the location where it began to spread to another place in the body is termed as metastatic cancer. Treatment can help to extend the lives of certain patients with metastatic cancer. In general, however, the main aim of metastatic cancer therapy is to monitor or reduce the effects of cancer growth. Metastatic cancers can cause significant harm to the functioning of the body and are the leading cause of deaths. However, depending on the type of cancer, several health-related complications can arise even for non-metastatic or primary cancers, some of which can be life threatening. Further, there's always a risk for metastatic growth due to the unpredictable nature of the cancer progression. Thus, clinical interference is inevitable for a cancer patient despite the intensity/extent of cancer.

1.2 Clinical management in Cancer

Cancer is one of the leading causes of death in the world. Cancer prevalence has been observed to increase dramatically with age, meaning that three of every 100 individuals develop cancer every year above the age of 60. In the last 30 years, the incidence of cancer cases has grown by about one-third. Latest figures show that the number of cases is already increasing at a rate of almost 1.5 per cent per year. It is estimated that proportion of the cancer-population over 65 would rise from 16% in 2004 to 23% by 2030, further escalating the total incidence (Bray and Møller, 2006).

The primary methods of treating cancer for many years have been surgery and radiotherapy. Controlling the primary tumour is a problem since it is responsible for major symptoms and health depletion of the patient. Also, failure to control the disease locally means certain death. As discussed in previous section, the most important cause of death is metastatic spread, which is the extension of the tumour of the cancer to other body parts. Therefore, early detection of cancer and treatment is vital. Once metastasized, it is almost difficult to treat the cancer patient. The prognosis in that case is not changed by the treatment of the primary tumour, although the symptoms can be mitigated.

Cancer is a chronic illness, and much like all other chronic medical conditions, cancer patients have families, careers and other obligations. The target of modern healthcare system, therefore, is to cure cancer if possible and, if not curable, to manage symptoms in order to increase the quality of life and to extend the life of the person by a significant time. For example in elderly population which are more prone to risk of death, a more careful and precise care system is required. Cancer is continually handled in a multidisciplinary team environment to boost the result and reduce the morbidity. Some centres make recommendations on multidisciplinary tumour committees, and some centres have specialized multidisciplinary facilities. Multidisciplinary team members cover doctors, radiation and surgical oncologists/hematologists, palliative care specialists, radiologists, pathologists, general physicians, nurses and allied health professionals. While, the patient management in cancer is a complicated and cancer specific process, it can be categorized into three broad steps viz. screening and diagnosis, evaluation of tumor extent and risk, and treatment. These steps are explained in the following sections.

1.3 Screening and diagnosis

The individuals which experience certain symptoms or difficulties along-with individuals who are at higher risk are often advised to undergo routine screening tests for different cancers. The goal of cancer screening is to allow the diagnosis earlier and thereby improve the survival rate. Highly responsive tests are important if the condition is curable at an early stage and if the effects of a false-positive result are not medically or mentally significant for patients. To cite some examples, smear test is a sensitive examination for cervical cancer and the diagnosis of cervical cancer can be readily confirmed by biopsy. Ultrasound and serum CA-125 screening is less sensitive but may prove to be of benefit for ovarian cancer. Breast, cervix and lung cancer are so prevalent at certain ages and in certain communities that screening is a realistic proposition. Clearly, the frequency of tumours must be high enough to warrant the screening programme. **Table 1.1** lists the symptoms and currently used screening/diagnostic tests of a few prevalent cancers.

Table 1.1 Some common symptoms and screening tests used for diagnosis of different cancers

Cancer-type	General symptoms	Common diagnostic tests
Breast cancer	lump, blood discharge from the nipple, change in shape or texture	mammography
Cervical cancer	bleeding, foul vaginal discharge, lower back or abdominal pain	cervical smear
Colorectal cancer	abdominal pain, blood in stool, change in bowel habits, stool inconsistency	faecal occult blood testing, rectal exam, flexible sigmoidoscopy
Lung cancer	cough with blood, chest pain, weight loss	chest radiography, chest CT scan
Ovarian and Uterine cancers	abdominal pain, bloating, loss of appetite	pelvic ultrasound, CA-125
Skin cancer	unusual growth or change in a mole	self-examination
Gastric cancer	bloating, nausea, heartburn, indigestion	radiological and endoscopic examination
Prostate cancer	difficulty in urination,	prostate-specific antigen (PSA)

These initial screening tests, if positive, are followed by biopsies for further confirmation of cancer. There are very few situations under which the diagnosis of tumour is rendered in the absence of pathological validation, especially because screening tests have gotten less invasive

over the last few years. A clinical diagnosis exclusively (no biopsy) is most commonly made in the case of severe/advanced disease in a low performance patient, where anti-cancer treatment does not increase quality of life or longevity. Thus, the major proportion of patients are diagnosed with cancer confirmed by tissue pathology. For this, the most important point is the retrieval of tissue sample through the least invasive method. For example fine needle aspiration biopsy (FNAB) for examination of lymph nodes in patients with lung or abdominal mass. These least invasive techniques such as FNAB or core biopsy enable effective staging and treatment strategy. The second-most vital point which is specifically paid attention to is the amount of tissue which is collected from the patient.

1.4 Risk evaluation and prognosis

After the cancer is diagnosed, the tumour extent and related risk is estimated. While the general prognosis of malignant tumours is most often summarised by showing the percentage of patients surviving at 5 or 10 years of age, these estimates generally disguise a wide variance in survival, spanning from treatment to demise within a few months of diagnosis. The hunt for prognostic markers has drawn the interest of oncologists for several years. The aim is to classify those patients for which a treatment plan is likely to be effective (for example surgery) and, likewise, those with whom it is supposed to fail, normally due to tumour spread past the primary site that is evident or visible. A new approach needs to be taken for these patients. The following sections elaborate various techniques which are currently used and/or are topics of active research.

1.4.1 Cancer staging: The TNM system

One component of the evaluation of factors which affect prognosis of any particular patient is the staging of the magnitude of the disease at diagnosis. The main features of careful tumour staging are: (i) it should be a standardized way to document primary tumour characteristics (ii) it should provide an efficient prognostic estimate (iii) it should complement the biological understanding of tumor and (iv) it should assist in efficient treatment design and planning for a patient. Out of the various staging systems available, the TNM staging system of American Joint Committee on Cancer (AJCC) is the most widely used for solid tumours such as breast-

Table 1.2 Cancer staging systems excluding AJCC TNM.

Staging system	Details
Ann Arbor	Lymphomas; uses roman numerals I-IV and E,S
Duke's	Colon cancer; similar to TNM, A-D for tumor spread, B & C have further divisions
WHO-CNS	Central nervous system; uses histological features
FIGO	Gynecological cancers; similar to TNM, Stage 0 doesn't exist
Cotsworld	Lymphoma; modifications to Ann Arbor, Stage III with further divisions
CIN	Cervical intraepithelial neoplasia; staged in grades (I-III), caused by HPV

-cancer, head and neck cancers and lung cancer. The AJCC's TNM system is simple to grasp and draws attention due to its relevance in cancer prognosis. However, it has certain limitations in some cancers such as leukaemias and lymphomas, for which specialized systems exist. **Table 1.2** mentions various staging systems other than AJCC's TNM.

There are three major biological components involved in the AJCC TNM staging i.e. extent and size of localized primary tumour (T), spread to nearby lymph nodes (N) and distant metastasis (M). Each of these have further divisions, generally notated using numeric or alphabetical suffixes (such as T1a, M0 etc.). Although, the definitions vary according to the type of cancer, it broadly follows the rules given in **Table 1.3**.

Table 1.3 The TNM staging rules.

T Stage		N Stage		M Stage	
x	-cannot be assessed	**x**	-cannot be assessed	**0**	-no metastasis
0	-no evidence of primary tumour	**0**	-no nodes	**1**	-distant metastasis observed
is	-in situ or localized tumour	**1-3**	-increasing number implies increasing number of nodes and/or size of nodes		
1-4	-increasing number implies bigger size and degree of invasion in the organ				

After the TNM assessment a Stage (0-IV) is assigned, wherein increasing number is a representative of severity of the associated cancer with Stage IV being the deadliest and often incurable cancer stage. **Figure 1.2** explains the staging process in melanoma (skin cancer) based on the tumour spread into the skin layers and beyond. The stages are often divided into further substages. A detailed staging process involving sub-staging in colon cancer is illustrated in **Figure 1.3**.

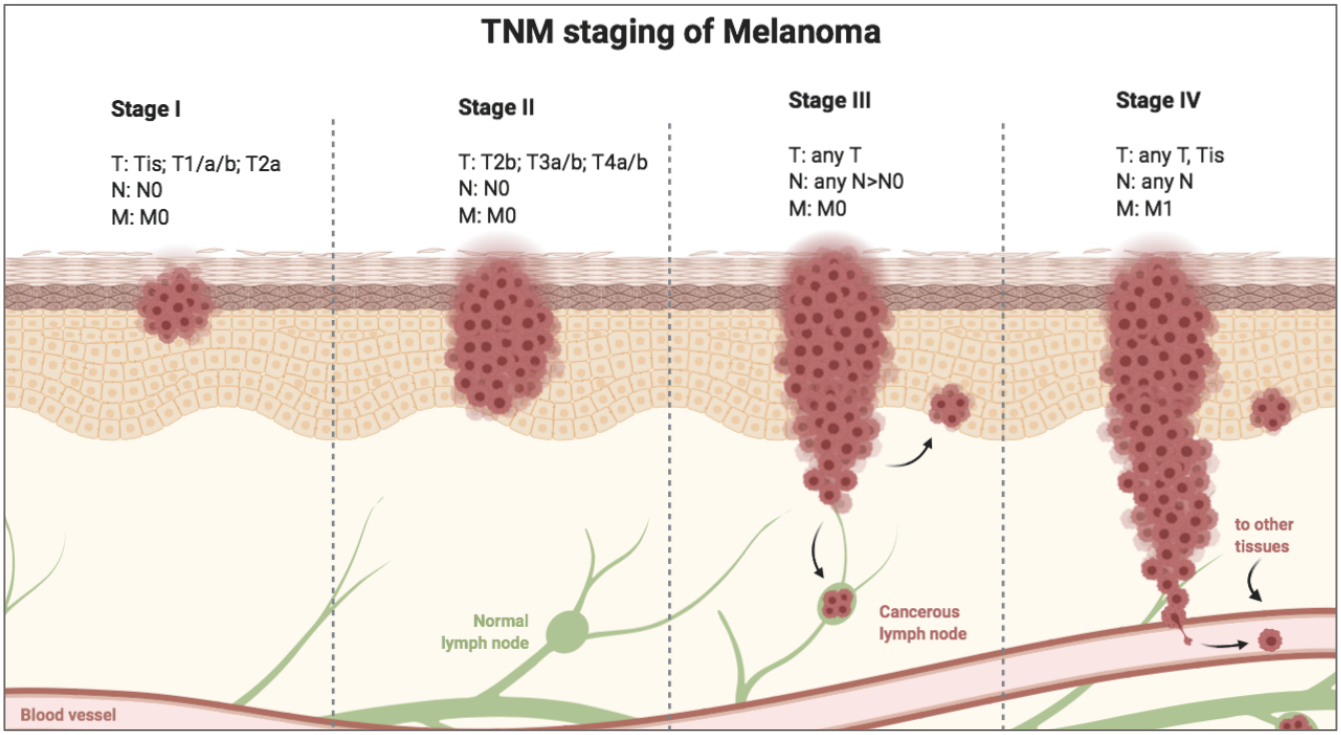


Figure 1.2 The AJCC TNM staging of melanoma of the skin. Here, a: without ulceration, b: with ulceration. (Source: biorender.com)

1.4.2 Non-anatomical prognostic factors

Apart from the information obtained from analyzing cancer spread associated parameters such as the ones involved TNM staging, various other factors are often considered in risk evaluation of cancer patients. These factors can be inclusive of clinical factors such as age, heritable genetic traits, genomic factors such as expression or mutation status of certain genes, concentration levels of certain proteins in blood or serum, environmental factors such as exposure to radiation to lifestyle related factors such as diet and smoking/drinking habit.

Previous experiments have shown that cancer cells release certain substances in the blood, known as tumor markers, which are often used at a preliminary screening level. Another group of substances known as tumour biomarkers, on the other hand, are not directly expressed by tumour cells but show an altered level when compared with a normal body. These altered levels may indicate presence of tumour and are used at both diagnostic and prognostic levels. Some commonly used markers in cancer are:

- **Prostate-specific antigen (PSA)**: PSA is a protein in the blood whose levels are increased in prostate cancer. PSA levels are used, to determine how a patient has responded to therapy. They are also used to scan for recurrence of the tumour. However, only PSA test cannot confirm the prostate cancer diagnosis.

- **Alpha-fetoprotein (AFP)**: AFP levels can be measured via blood tests. A high amount of AFP is suggestive of liver cancer or germ cell tumors of the testes or ovaries. However, elevated AFP levels are also found in pregnant women and can also be caused by disorders such as chronic active hepatitis.

- **Human chorionic gonadotropin (HCG)**: An increase in hCG or b-hCG in blood is indicative of cancer in the liver, pancreas, testis, ovary, stomach and lung. It is also elevated during pregnancy and thus must be ruled out before.

- **Carcinoembryonic antigen (CEA)**: The most popular cancer in which this tumour indicator is used is colorectal cancer, although many other cancers, such as epithelial cancers, also exhibit an increase in its levels.

- **CA 125**: Cancer antigen 125 (CA 125) is a protein in blood which is mainly associated with ovarian cancer diagnosis and management. CA 125 levels have been observed to be elevated in other cancers such as uterine, pancreatic, colorectal, breast and cervix.

- **CA 19-9**: Cancer antigen 19-9 (CA 19-9) is a protein whose high levels are consistent with colon, lung, and bile duct cancers. Elevated levels of CA 19-9 can suggest advanced pancreatic cancer, but noncancerous disorders, including gallstones, pancreatitis, liver cirrhosis and cholecystitis, are also associated with it.

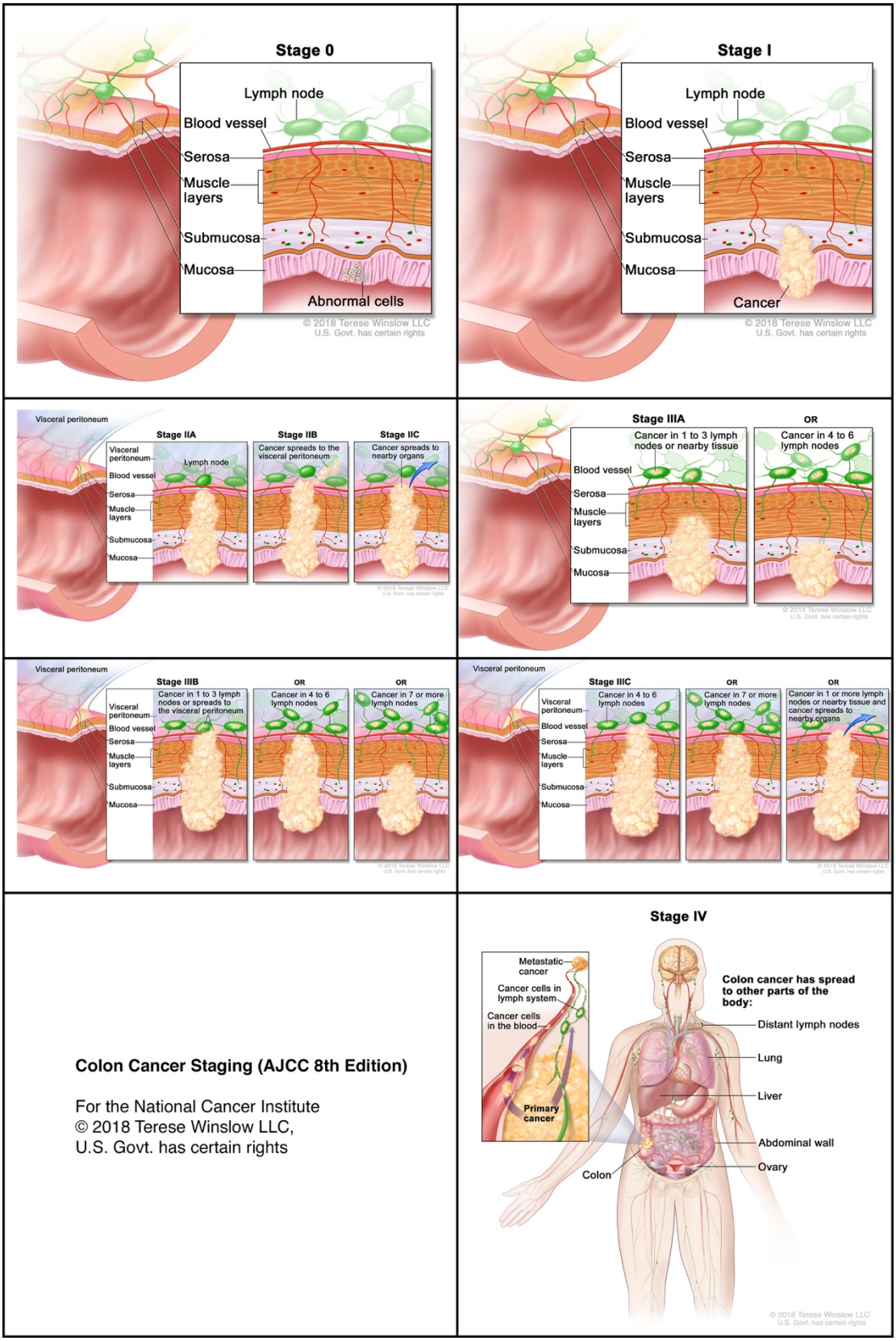


Figure 1.3 The AJCC TNM staging in colon cancer. (Permission to use. For the National Cancer Institute © 2018 Terese Winslow LLC, U.S. Govt. has certain rights)

Other markers which are prominently used for specific cancers include thyroglobulin (thyroid cancer), chromogranin-A (neuroendocrine carcinoma), neuron specific enolase (lung small cell carcinoma), lactate dehydrogenase (melanoma, lung cancer, germ cell cancers) and CA 15-3 (breast cancer). Several genetic biomarkers/tests for cancer prognosis have also been proposed and some of them are in active use at a clinical level such as OncotypeDx (breast cancer), DecisionDx (melanoma) and Oncodefender-CRC (colorectal cancer). Much of these markers/biomarkers are used as adjunct tools for diagnosis, prognosis and management in cancer patients and they are not often relied on blindly due to their limitations. As a result, only a handful of these are included in the AJCC staging along with other factors (**Table 1.4**). The hunt for efficient cancer biomarkers is an ongoing process and demands multidisciplinary efforts. Further, AJCC staging principles are regularly updated to include novel prognostic factors and provided in AJCC staging manual (Amin *et al.*, 2017).

Table 1.4 Non-anatomical prognostic factors included in AJCC staging (2018)

Cancer	Prognostic Factor	Test
Melanoma	LDH	Blood test
Prostate cancer	PSA Gleason Score	Blood test
Breast cancer	ER, PR, HER2 status OncotypeDx	Genetic test (biopsy) Genetic test (biopsy)
Testicular cancer	LDH, HCG, AFP	Blood (LDH/AFP/HCG) or urine test (HCG)
Gestational trophoblastic neoplasms	Risk Score	Clinical factors (Monitoring of patient)
Thyroid cancer	Age	Clinical factors (Monitoring of patient)
Retinoblastoma	Rb1 mutations	Genetic test (tissue sample)
Primary cutaneous lymphomas	Peripheral blood involvement	Peripheral blood smear test
Oropharyngeal cancer	p16/ HPV status	HPV test (tissue sample)
Gastrointestinal stromal tumour	Mitotic rate	Biopsy

1.5 Treatment

Similar to any other treatment, the main aim of cancer treatment is to cure cancer in diagnosed patients. Wherever, this is not possible such as in late stage patients, the treatment often involves

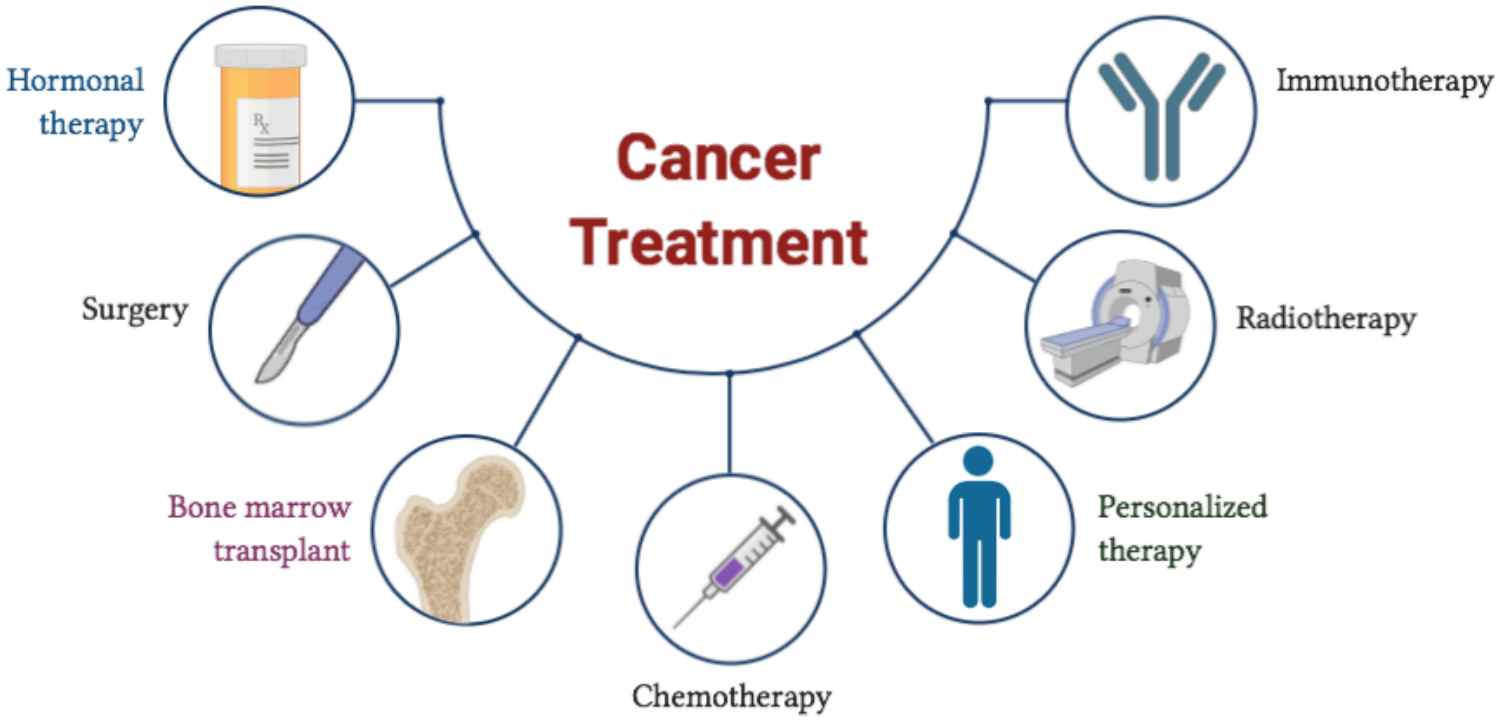


Figure 1.4 Types of cancer treatment

shrinking the cancer or deterring the cancer growth to allow the patients to have a longer life span which is symptom free. There are two major steps involved in cancer treatment: (i) Primary treatment which is focused on removal of primary tumour and/or killing of all the cancer cells. Surgery is the most common primary treatment, however, depending on the sensitivity of cancer cells to chemo or radiotherapy, those treatments may also be used. (ii) Adjuvant therapy wherein the main goal is to kill cancer cells that remain after primary treatment, so as to avoid cancer recurrence. Most commonly used methods involve chemotherapy, radiotherapy and hormone therapy. **Figure 1.4** mentions the treatment options currently in practice and are explained in details below:

- **Surgery-** The main aim of cancer surgery is to heal your cancer by extracting all of it from the body. Typically, the surgeon does this by cutting and removing the cancerous tissue

while leaving the adjacent healthy tissue unaltered. In order to decide if the cancer has spread, the surgeon can even remove several lymph nodes in the region. This lets the doctor determine the likelihood that you will be healed, as well as the need for further care. For example mastectomy is the removal of a whole breast in the case of breast cancer surgery while lumpectomy is removal of a part of breast. Similarly, in lung cancer surgery, lobectomy is removal of a part of lung and pneumonectomy is the removal of whole lung. Surgery is often combined with other treatments such as chemo or radiotherapy.

- **Hormonal Therapy-** Hormone therapy or hormonal therapy or endocrine therapy is a cancer therapy that delays or prevents the development of cancers that use hormones for growth. Hormone treatment can be divided into two broad classes, those that inhibit the body's capacity to generate hormones and those that meddle with the function of hormones in the body. Hormone therapy is most widely used to treat prostate and breast tumours that use growth hormones. It is also commonly used along with other cancer therapies.

- **Bone Marrow Transplant-** For blood cancers such as lymphomas and leukemias, the target for both chemotherapy and radio therapy is often bone marrow. As a result, stem cells are damaged which causes a fall in production of healthy blood cells and leads to several health issues. To deal with this, stem cells are transplanted from a healthy donor. Donor stem cells restore the bone marrow and blood cell production.

- **Chemotherapy-** Chemotherapy is a method of cancer treatment that involves drugs to destroy cancer cells. Chemotherapy operates by preventing or slowing down the progression of cancer cells that expand and multiply rapidly. Chemotherapy is used to treat cancer, minimize the risk that it will return or interrupt or slow down its growth. Chemotherapy is also used to reduce tumours that cause discomfort and other complications. Chemotherapy is performed in various ways, including oral, intravenous, infusion, intra-arterial, etc.

- **Radiotherapy-** Radiation or Radio therapy is a method of cancer treatment that destroys cancer cells using beams of strong radiation. By disrupting the genetic material that

regulates how cells expand and differentiate, radiotherapy destroys cells. Although radiotherapy kills both healthy and cancerous cells, its purpose is to kill as few normal/healthy cells as possible. Much of the damage caused by radiation will also be healed by normal cells. The treatment is distributed over several weeks to allow this healing process. Examples of cancers sensitive to radiotherapy include lymphoma and seminoma of testis.

- **Immunotherapy-** Immunotherapy is a form of cancer treatment that improves the body's natural defenses against cancer. It uses chemicals made by the body to enhance the way the immune system operates to detect and kill cancer cells. The immune system consists of a complicated mechanism used by the body to combat disease. This mechanism includes cells, organs, and proteins. Cancer can usually bypass many of the normal defences of the immune system, enabling cancerous cells to continue growing. Various forms of immunotherapies function in various ways. Some of these help the immune system to stop or slow the growth of cancer cells. Others help the immune system to destroy cancer cells or stop cancer from spreading to other parts of the body. Immunotherapy can be used alone or in conjunction with other cancer treatments. There are several forms of immunotherapy including monoclonal antibodies and tumour agnostic therapies (such as control point inhibitors), oncolytic viral treatment, T-cell therapy, and cancer vaccines.

- **Personalized Therapy-** Personalized therapy or Precision therapy is a means for health care providers to provide and prepare personalised care for their patients, depending on the genes of the individual or the genes in their cancer cells. Precision therapy explores how a certain gene alteration/mutation could influence a person's likelihood of having cancer and how their genes could affect treatment. The method incorporates knowledge from genetic testing to help clinicians establish a treatment strategy that typically contains very detailed guidelines. Precision medicine can help allow a more precise diagnosis and improve recovery in some cases. In other cases, it can encourage people to make choices about healthier behaviours, early screening tests, and other protective measures whether they are at risk for a specific cancer.

1.6 Apoptosis in cancer: biological and therapeutic role

As discussed above, cancer cells acquire certain capabilities through which they grow and survive. Out of the major hallmarks of cancer, evasion of apoptosis is one of prominent hallmark. A large

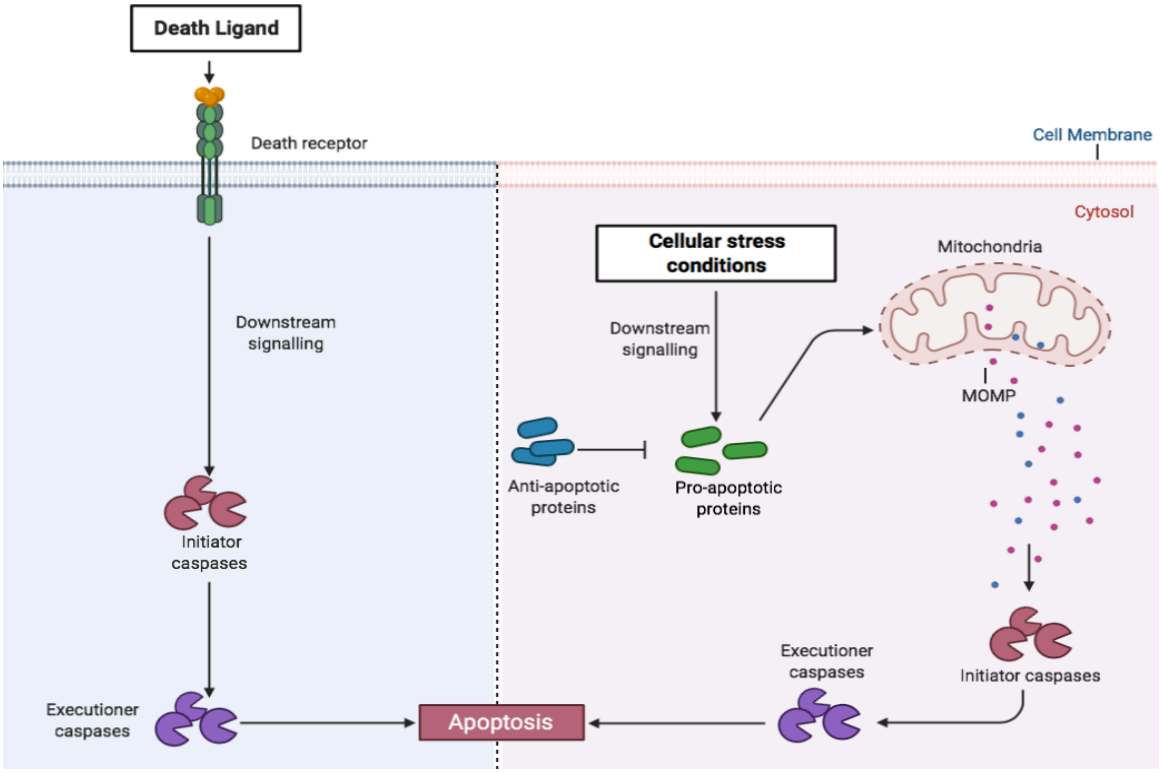


Figure 1.5 The pathways involved in apoptosis. (source: biorender.com)

number of recent advances in oncology are focused on developing chemotherapeutic drugs that kill cancer cells via induction of apoptosis. Apoptosis is a precisely regulated cell death event with distinctive genetic and biochemical mechanisms that play a crucial role in the growth and homeostasis of normal tissues. It leads to removing excessive and undesirable cells in order to maintain a healthy equilibrium between cell viability and cell depletion in organisms. Previous studies suggest that inadequate apoptosis can propagate as cancer or autoimmune disorders. Dysfunction of the apoptotic pathway not only facilitates carcinogenesis, but also makes cancer cells immune to treatment.

Apoptotic cells show peculiar features during the apoptotic process. Usually, the cell starts to shrink after the cleavage of laminates and actin filaments in the cytoskeleton. The apoptotic degradation of chromatin in the nucleus frequently contributes to nuclear condensation. Cells further keep shrinking, packing themselves in a manner that requires macrophages to eliminate

them. These phagocytic cells (macrophages) are responsible for removing apoptotic cells from tissues in a neat and orderly way. In order to facilitate their phagocytosis, apoptotic cells also undergo changes in the plasma membrane that cause a macrophage response. One such modification is the translocation of phosphatidylserine from the interior of the cell to the external surface. The final stages of apoptosis are also indicated by the formation of membrane blebs or blisters.

The key players of apoptosis are caspases, since they act as both executioners as well as initiators of apoptotic process. There are two major pathways that initiate the caspases: the intrinsic or mitochondrial pathway and the extrinsic pathway. Both of these lead to a common pathway involving executioner caspases. This is followed by cleavage of caspase-activated deoxyribonuclease inhibitor that is critical for nuclear apoptosis. In conjunction, downstream caspases cause cleavage of DNA repair proteins, protein kinases, inhibitory endonuclease subunits and cytoskeletal proteins. They also have an impact on the cell cycle related signaling pathways and cytoskeleton, which together lead to the expected morphological variations in cell death. The intrinsic pathway of apoptosis is a programmed mechanism within the cell through which various stresses are dealt with for example hypoxia, DNA damage that cannot be repaired, high oxidative stress etc. The apoptotic outcome in this pathway depends on the integrity of mitochondrial membrane, which, when disrupted results in the release of molecules which activate executioner caspases. The regulators of this pathway are mainly proteins of Bcl2 family which are divided into two groups i.e. pro-apoptotic and anti-apoptotic molecules. As the name implies, pro-apoptotic molecules cause the increase in mitochondrial membrane permeability whereas anti-apoptotic molecules obstruct this process. The extrinsic pathway on the other hand consists of specialized death receptors which are activated by certain (death) ligands. A subsequent activation of initiator caspases and thereafter executioner caspases then leads to cell demise. The mechanism is demonstrated in **Figure 1.5** and a detailed explanation is provided in the next chapter.

1.7 Origin of proposal and thesis objectives

A lot of efforts have been invested in the past decades, to maneuver the apoptotic machinery in an anti-cancer direction. Several key regulators and their roles in this complicated mechanism have been elucidated. Briefly, it has been observed that certain components and parts of the apoptotic process are compromised in cancer cells due to which these damaged cells refuse to die and

propagate the damage into further generations. This current understanding of the apoptotic pathways has led to the development of drugs which target these crucial elements and restore the survival/death balance. Additionally, the altered levels or status of the apoptotic regulators are also utilized for cancer prognosis and risk prediction. However, there still remains the challenge for development of novel prognostic biomarkers/methods for risk evaluation of cancer patients. Additionally, due to the relevance of various clinical factors in cancer development and growth, these upcoming methods should integrate relevant features in order to complement the existing risk prediction systems or replace them entirely. The novel prognostic methods can be utilized for more accurate risk estimation and thereby effective therapeutic planning.

In order to augment the knowledge regarding the role of apoptotic pathway in conjunction with clinical factors in cancer prognosis, the current studies aim to evaluate the prognostic strength of various apoptotic genes/proteins. This information is further used to develop models which could be utilized for risk evaluation in different cancers. Wherever possible, a comprehensive contrast is made between the clinical factors and expression-based models. The resultant models thus involve the most relevant features only. The *in-silico* models proposed in the study are intended to bear the following major characteristics: (a) They should be minimally invasive, (b) They should be cost effective, (c) They should be universal or applicable across many cancer types, (d) They utilize recent data and (e) They are serviceable to community in the form of free web-based service or mobile app. The study is broadly categorized according to the following objectives:

(a) Development of proteomic data based prognostic models
(b) Development of genomic data based prognostic models
(c) Development of clinical factors-based prognostic models
(d) Development of universal prognostic models

1.8 Thesis organization

The specific aim of this thesis involves a comprehensive analysis of apoptosis related molecular data (protein/gene expression) in the context of cancer prognosis. However, since traditional methods employ clinical data for risk prediction/staging in cancer, one cannot simply ignore their relevance. To address both these issues, our study delves into delineating the prognostic ability of both clincal and molecular data for various cancers. We specifically address three major cancer-

types (i) Colorectal cancer: To illuminate the predictive strength of apopotosis related proteins in therapeutic decision making; (ii) Thyroid Cancer: The superiority of gene expression based prognostic models as compared to traditional methods; (iii) Melanoma: The conflicting case wherein expression based methods fail and clinical data based approach triumphs and (iv) Universal prognostic biomarker: Establishing a generic biomarker which can be applicable across a range of cancers. Overall, This thesis is organized into eight chapters containing the information as explained below:

Chapter 1- Introduction to cancer and the fundamental biological understanding of the condition is presented. This is followed by the brief information about the clinical management pipeline currently used in the healthcare system and the relevance of risk evaluation procedures involved. Finally, the role of apoptotic pathway in carcinogenesis and cancer treatment is discussed. The end of this chapter emphasizes the need to study apoptotic mechanism in contrast to other c linical factors for identification of novel prognostic biomarkers and development of efficient risk prediction models.

Chapter 2- This chapter presents a literature survey regarding the cancer biomarkers and the relevance of omics-based biomarkers in cancer management. It also highlights the importance of various clinical 'risk' factors associated with cancer. The apoptotic pathways are presented in detail and discussed in the context of biomarker discovery and treatment. Briefly, this chapter lays down the motivation behind the study.

Chapter 3- Apoptosis related proteins have been widely associated with the prognosis of colorectal cancer in the past. This chapter covers a study, wherein, proteomic data related to intrinsic mitochondrial pathway proteins was utilized to develop a novel predictive biomarker for colorectal cancer patients. The proposed biomarker is evaluated against several clinical factors and a previously established biomarker. Ultimately, a web-based tool, which encompasses the proposed biomarker, is presented for risk prediction in colorectal cancer patients.

Chapter 4- At the gene regulatory level, the mechanism of apoptosis is a sophisticated multi-level process involving a wide quantity of genes. Some of these genes are also part of other mechanistic

pathways. Since the apoptotic defect can arise at any of the several regulatory steps in this process, in this chapter, the genomic data corresponding to the complete apoptotic regulatory pathway is employed. Thereafter the prognostic relevance of each of these genes is analyzed in the context of thyroid cancer. Key genes are identified and validated through their published roles in thyroid cancer. Other validation studies such as differential expression amongst tumor and normal tissues as well as protein level expressions are also studied. The importance of clinical features and the benefit of integrating 'age' in the proposed genetic biomarker is also discussed.

Chapter 5- In this chapter, a comparative analysis amongst the various cancer related pathways, including apoptosis, and clinical factors is investigated through the lens of "prognostic value" for the case of melanoma. Developed models are assessed for their predictive ability and, subsequently, the superior significance of clinical factors in melanoma prognosis is established. A novel risk grading method is proposed and compared against a popular tool. Ultimately, the usage and functionality of the web-resource and mobile application, which implements the proposed prognostic model is presented.

Chapter 6- The major aim of this chapter is to extend the concept of chapter 4 to other cancers and utilize the information for developing universal prognostic models. A universal prognostic biomarker which is applicable across a variety of cancers can have a huge implication for the future. The chapter discusses the development of a 11-gene based biomarker and further proposes a strategy for development of cross-cancer biomarkers.

Chapter 7- The penultimate chapter of this thesis discusses the role of various extrinsic and intrinsic risk factors in Cancer. Due to their major contribution in cancer risk and mortality, clinical features such as lifestyle related habits including smoking/drinking, age, gender, race, heritable factors, environmental exposure etc. need to be further explored in conjunction to existing staging system. However, due to the emergence of "omics" based biomarkers, these have been largely undermined and are a matter of consistent debate. This chapter focusses on mining the prognostic strength of various clinical factors and offers novel models for risk assessment in multiple cancers.

Chapter 8- This chapter concludes the thesis work with a brief outline of the work and its contribution to the field of Cancer.

Github Repository

A Github repository (https://github.com/raghavagps/Chakit_Thesis) is also provided with the necessary datasets and scripts.

2

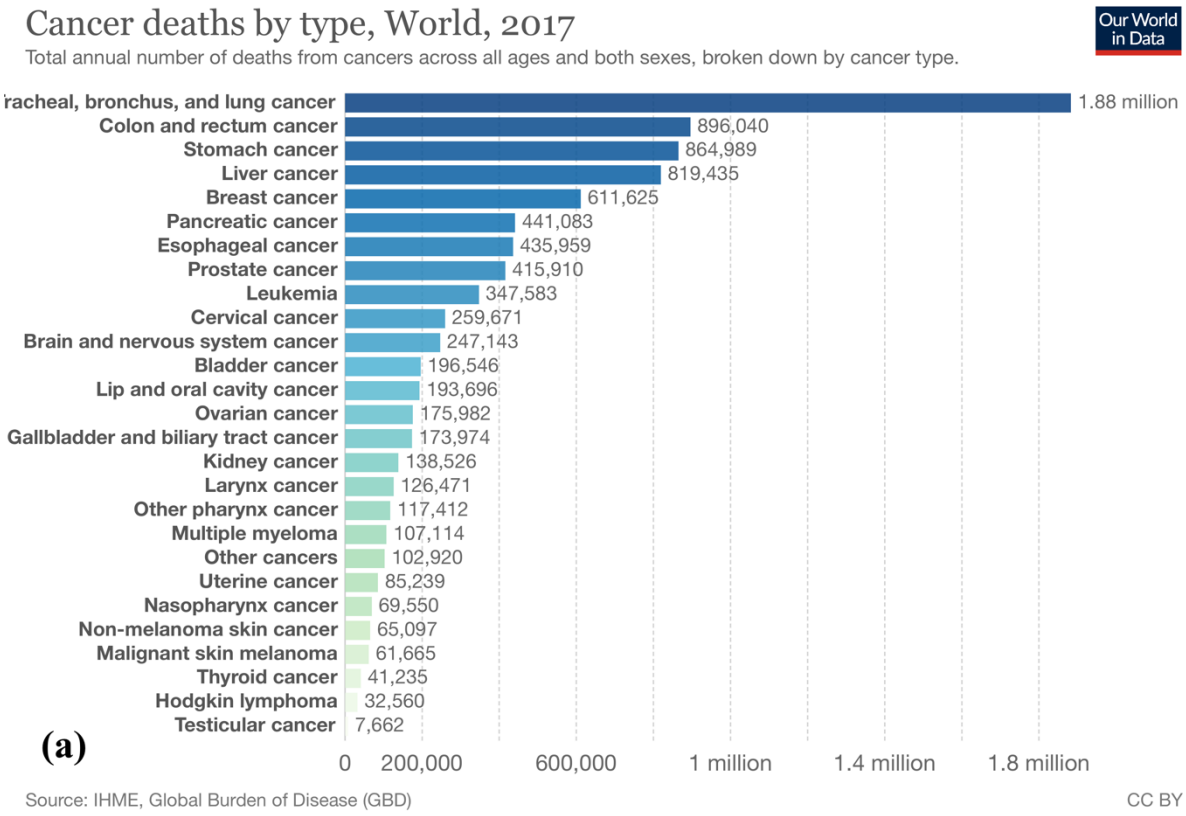
Review of Literature

2.1 Cancer: A global nuisance

Cancer is the second leading cause of deaths worldwide, with around one in every six deaths in 2018. In United States alone, 1806590 new cancer cases and 606520 deaths have been projected for the year 2020, in a latest report by American Cancer Society (Siegel *et al.*, 2020). According to this estimation, Lung cancer is the major cause of deaths in both the sexes (23% in males and 22% in females) while 10% of all the cancer related deaths in Males is due to Prostate cancer and 15% of all the cancer related deaths in females is due to Breast cancer. Colo-rectal cancer is now the third leading cause of death in both the sexes followed by other cancers as opposed to its second rank in 2017 (**Figure 2.1a**), due to implementation of strict screening and prognostic procedures. The National Cancer Registry Programme report for the year 2020 also estimated a whopping 1392179 cancer incidences in India (Mathur *et al.*, 2020). While a lot of clinical efforts and ground-breaking research has been responsible for the 29% overall decline of cancer deaths from the peak cancer death rate of 1991, the rates are still increasing, however, with a slower momentum. **Figure 2.1b** shows the trend of increasing death rates after 1990. As per the data, the concern is very serious since a cancer mortality-free future is still unforeseeable. The WHO-GLOBOCAN database states that, owing to increased socio-economic growth, rates and types of cancer cases from developed countries are increasingly moving towards developing countries. Additionally, due to their geographical spread, there are certain different cancer forms and local risk factors for each region, such as dietary habits and environmental exposure, which also play a major role in the occurrence of new cancer cases. From these points, it is quite clear that cancer is a global burden, and there is an immediate need to design solutions to treat this disease to boost patients' life expectancy.

2.2 Factors associated with risks of cancer

Cancer is unpredictable and as a result, typically, it is not possible to determine precisely why one individual gets cancer and another person does not. But various studies and statistical tests have found that certain risk factors (hereby used collectively in the term '*clinical factors/features'*) strengthen the likelihood of an individual getting cancer. However, there are also factors that are associated with a lower cancer risk which are often referred to as protective risk factors or protective factors.



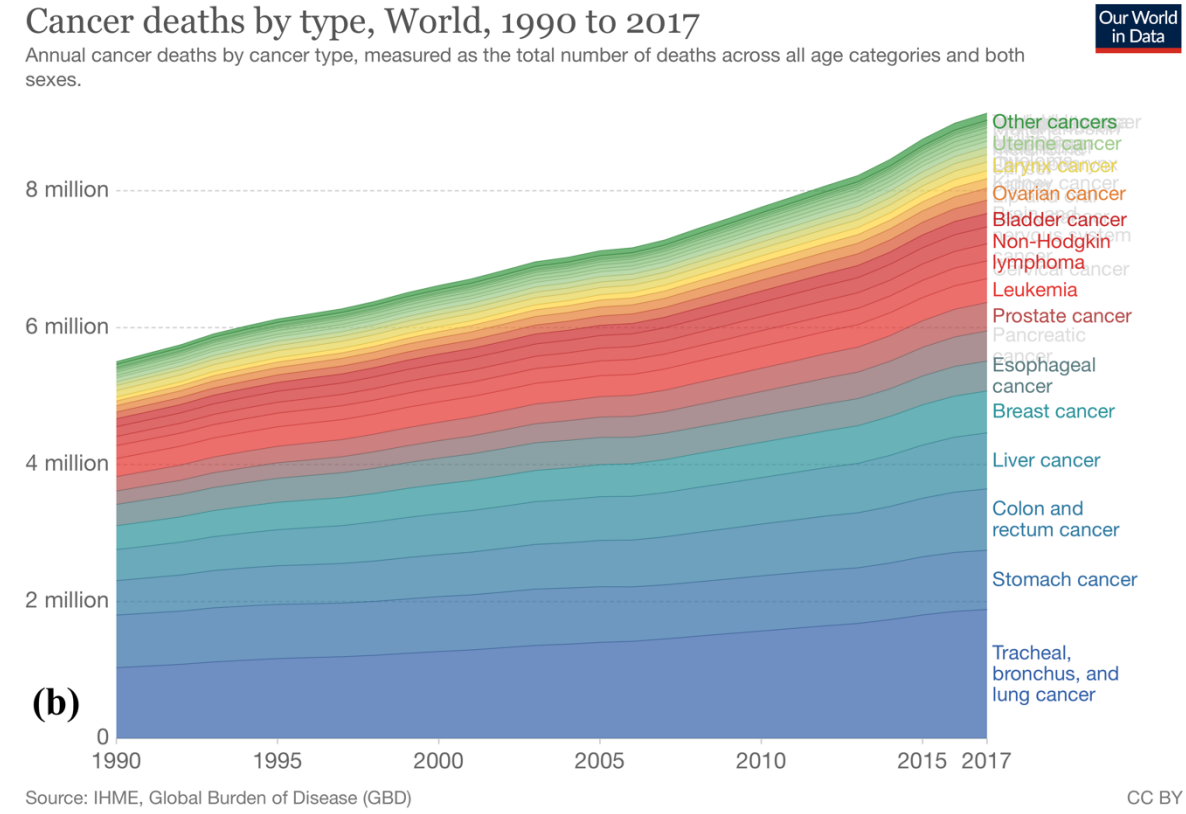


Figure 2.1 The global mortality burden of Cancer. (a) Distribution by cancer types. (b) The death rate trend of various cancers. (source: http://ourworldindata.org/)

Cancer risk factors include proximity, as well as certain habits, to toxins or other compounds. They also involve factors, such as age and family history of cancers. Broadly the factors associated with risk of getting cancer or intensification of cancer can be divided into four major categories as shown in **Figure 2.2** and explained below:

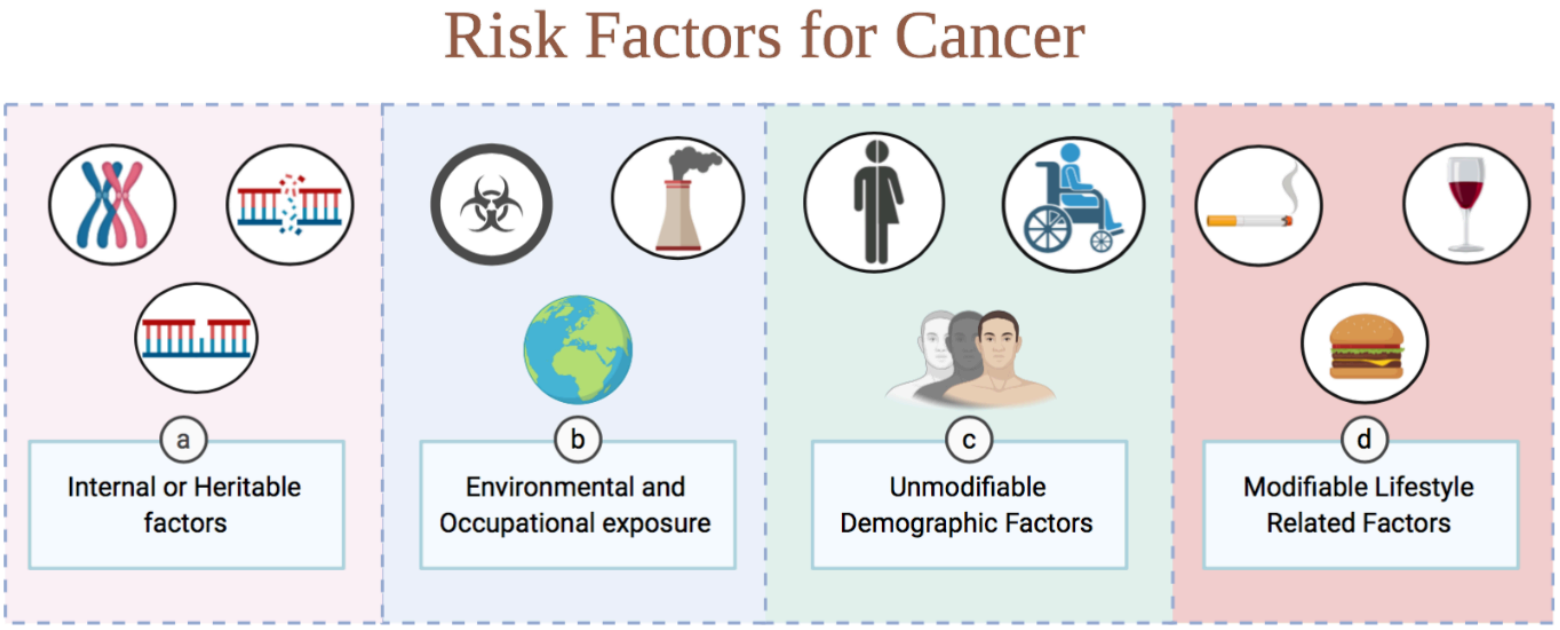


Figure 2.2 The risk factors associated with Cancer.

2.2.1 Internal or Heritable factors

Some forms of cancer tend to occur in some families. This could be because family members share some habits or conditions that raise the risk of cancer. Just about 5% to 10% of all cancers are directly caused by hereditary mutations inherited by the parent. Most family cancer syndromes are caused by hereditary mutations in tumour suppression genes. The most common examples of heritable cancers are HBOC or hereditary breast or ovarian cancer, Lynch syndrome and Li-Fraumeni syndrome. HBOC is caused by inherited mutations in BRCA1 or BRCA2 genes (Grill *et al.*, 2020; Hodgson and Turashvili, 2020; Yoshida, 2020). In addition to breast and ovarian cancer, this mutation may also lead to fallopian tube cancer, pancreatic cancer, primary peritoneal cancer, prostate cancer and male breast cancer, as well as other cancers. The most common hereditary syndrome that raises the risk of colon cancer in an individual is Lynch syndrome. It is caused by mutation in any of several genes for mismatch repair (MMR) such as MLH1, MSH6, PMS2 etc. (Sinicrope, 2018; Lynch *et al.*, 2015). Li-Fraumeni is a rare inherited syndrome which

can cause a number of cancers to develop. Inherited mutations in the TP53 gene, which is a tumour suppressor gene and CHEK2 gene which is involved in DNA damage repair, are most frequent in this disorder (Grill *et al.*, 2020; Guha and Malkin, 2017; Correa, 2016). In their lifetime, individuals with this syndrome can develop more than one cancer. They also seem to have a higher risk from radiation therapy of getting cancer.

2.2.2 Environmental or occupational exposure

The environmental exposure basically involves exposures to radiation and other chemicals. Some major contributors to environment related cancer risk are air and water pollution, ionizing radiation such as X-rays and non-ionizing radiation such as UV rays. Air pollution has been claimed to increase the risk of lung cancer. The water treatment involves chlorination which is known to generate mutation-causing substances and thereby increase the risk of cancer for example genitourinary cancer (Koivusalo and Vartiainen, 1997; Mughal, 1992). The presence of arsenic in ground water is also presents a comparable risk (Christoforidou *et al.*, 2013; Smith *et al.*, 1992). It has been documented that ionizing radiation induces 1-3 percent of all cancers. In fact, radiation greatly raises the risk of leukemia, as well as breast, thyroid, bladder and lung cancer. In fact, the cancer patients treated with radiation are also at high risk of developing cancers and therefore radiotherapy is always assessed carefully. Ultraviolet radiation, and also magnetic and electrical fields, are involved in non-ionizing radiation. UV is mainly absorbed from sun-rays which induces cancers of the skin (Watson *et al.*, 2016; Narayanan *et al.*, 2010). Prolonged burning of the skin because of too much UV radiation, notably in childhood and youth, is the main cause of cutaneous melanoma. Fair-skinned, blue-eyed individuals with easily burnt and badly tanned skin are especially at risk. Occupational or work-related exposure mainly consists of individuals who are exposed to carcinogens, have a sedentary or low physical activity work life or are relatively more exposed to sun such as fishermen/seamen (Brown *et al.*, 2012; Kerr *et al.*, 2017).

2.2.3 Unmodifiable demographic factors

The unmodifiable risk factors of cancer contain factors like age, gender, ethnicity and socio-economic level. As the name implies, these factors cannot be controlled or modified. Aging is linked with most cancers. The longer a human lives, the more likely it is that their cells will accumulate alterations that causes cancer. With age, cells' ability to resist and rebound from these

defects weakens. Age for example is included in the AJCC staging of thyroid cancer and is believed to a significant factor. There are substantial differences between male and female in the rates of cancers other than gender-only cancers (such as breast cancer and prostate cancer) (Kim *et al.*, 2015; Donington and Colson, 2011). For example the age-adjusted laryngeal cancer morbidity in males, for example, is almost ten-fold relative to females. Melanoma is a prime example of ethnicity/race associated cancer, since skin color is directly associated with melanin and thereby melanocytic tumours (Gloster and Neal, 2006).

2.2.4 Modifiable lifestyle factors

Unlike the demographic factors, the lifestyle related factors such as diet, living habits, physical activity etc. are completely under control of the individuals. The single most significant factor in rising cancer risk is the consumption of tobacco products (Sasco *et al.*, 2004; Loeb *et al.*, 1984; Samet, 2013). The risk of contracting lung cancer is directly correlated with the age people start smoking along-with the daily amount. The consumption of heavy drinking also raises cancer risk and induces some apparent health issues (Boffetta and Hashibe, 2006; Braillon, 2018). It is considered that diet has the greatest effects on the risk of gastric, breast and lung cancer (Kerr *et al.*, 2017; Grosso *et al.*, 2017; Key *et al.*, 2020). Cancer tissue requires energy and minerals, so diet can have an effect not only on cancer formation, but also on its development. The risk of cancer can be raised by consuming processed meat. Numerous studies have found the link between physical exercise and the prevention of cancer (Kerr *et al.*, 2017; Brown *et al.*, 2012). Scientific evidence has accrued that physical exercise is especially protective against cancer of the breast, prostate, endometrium and colon. The cancer preventive impact of physical activity can be enhanced by fast exercise multiple days a week.

2.3 Cancer Biomarkers

Cancer biomarkers are biological molecules made in response to the tumour by either the tumour cells or any other body cells. These molecules are often used in risk estimation, as a diagnostic, prognostic or predictive indicator of a patient's outcome. **Table 2.1** lists the FDA approved biomarkers for various cancers. The subpopulations of patients which are most likely to respond to a given therapy can also be identified by cancer biomarkers (Goossens *et al.*, 2015). Biomarkers

may include chromosomes, gene products, particular cells, chemicals, enzymes, or hormones that are measurable in the blood, urine, tissues,

Table 2.1 The list of FDA approved cancer biomarkers

Biomarker	Cancer	Utility
Prostate-specific antigen (PSA)	Prostate cancer	Screening, Diagnosis
Carbohydrate antigen 125 (CA125)	Ovarian cancer	Diagnosis, Prognosis, Predictive
Carcinoembryonic antigen (CEA)	Colorectal/hepatic cancer	Prognosis, Predictive
Carbohydrate antigen 15.3 (CA 15-3)	Breast cancer	Predictive
Estrogen, progesterone receptors (ER and PR)	Breast cancer	Predictive (Hormonal therapy)
HER2	Breast cancer	Predictive (Traztuzumab therapy)
Carbohydrate antigen 27.29 (CA27.29)	Breast cancer	Predictive
Human chorionic gonadotropin-β (HCG-β)	Testicular cancer	Diagnosis, Staging, Predictive
Alfa-fetoprotein	Hepatocellular carcinoma	Diagnosis, Predictive
Calcitonin	Medullary thyroid carcinoma	Diagnosis, Predictive
Thyroglobulin	Thyroid cancer	Predictive
CA 19-9	Pancreatic cancer	Diagnosis
Nuclear matrix protein 22 (NMP-22)	Bladder cancer	Screening, Prognosis
Prostate cancer antigen 3 (PCA3)	Prostate cancer	Prognosis

or fluids of the body (Rhea and Molinaro, 2011). In certain patients with a particular form of cancer, genetic modifications in cancer cells, including point mutations, gene rearrangement or amplification, and resulting disruptions of cell division and proliferation are manifested by the release of biomarkers of those alterations. These can be used as biomarkers for the diagnosis of cancer or for forecasting reactions to different therapies (Sidransky, 2002; Vogelstein and Kinzler, 2004; Weissleder and Ntziachristos, 2003). Apart from screening biomarkers, which were covered in the last chapter, the major types of cancer biomarkers are explained below:

2.3.1 Diagnostic biomarkers

A diagnostic biomarker is used to identify individuals who have cancer. In comparison to a screening biomarker that would be applicable exclusively to symptomatic people, a diagnostic biomarker would be used only for asymptomatic cases. Interestingly, the properties of an optimal biomarker for diagnosis are identical to those for screening. Notably, most well-established

screening biomarkers could be used as diagnostic markers and PSA is a well-recognized example. The most widely used screening technique for prostate cancer is PSA paired with a digital rectal exam. Presently used cancer biomarkers have poor diagnostic sensitivity and specificity in contrast to the higher sensitivity expected from a good diagnostic biomarker (Pavlou *et al.*, 2013). For example, one of the best and most well-established diagnostic markers of multiple myeloma remains the Bence-Jones protein in urine (Kulasingam and Diamandis, 2008). However, some biomarkers have proven useful in verifying diagnosis, often in combination with other biomarkers. These have been used, in particular, to classify primary tumours with uncertain primary and/or other clinical imaging methods in metastatic cases. These combinations are known as biomarker panels or signatures (Henry and Hayes, 2012). For example Mor et al. (Mor *et al.*, 2005) stated in 2005 that a panel composed of four biomarkers (prolactin, osteopontin, leptin, and insulin-like growth factor 2) had a 95 percent sensitivity and a 95 percent specificity for ovarian cancer diagnosis jointly. In a further study, by addition of two more biomarkers to this panel (CA-125 and macrophage inhibitory factor), the specificity was shown to increase to 99.4% (Visintin *et al.*, 2008).

2.3.2 Prognostic biomarkers

Prognosis is the likelihood of any patient's treatment or possible future. At the time of diagnosis, a prognostic marker is a patient trait attribute irrespective of therapy; thus, a prognostic marker will provide details about the disease's likely outcome. The degree of increase in prognostic biomarker levels typically represents tumour burden, thus higher biomarker elevation reflects poor prognosis and vice versa. Prognostic biomarker can also be used in the staging system for cancer or the grouping of tumor-node-metastasis (TNM). For example very high levels of biomarkers such as AFP, LDH, and HCG-β can suggest advanced cancer with grim prognosis and result in testicular cancer, so that such biomarkers could be used for staging in the TNMS system with a site-specific prognostic factor (Szymendera *et al.*, 1981). A widely used prognostic and predictive biomarker is Estrogen receptor (ER) for breast cancer patients. ER positive patients have a good prognosis as well as respond to selective ER modulators and inhibitors. On the other hand, ER negative patients have a poor prognosis and also do not respond well to hormonal therapy (Duffy, 2005). In the same context, Progestrone receptor (PR) and HER2 levels are often used for their prognostic and predictive value in breast cancer patient management (Burstein *et al.*, 2001).

2.3.3 Predictive biomarkers

The response to numerous therapeutic procedures is often assessed by utilizing a predictive biomarker; hence, a predictive biomarker is the fundamental 'term' for personalised medicine (Pavlou *et al.*, 2013). Predictive biomarkers are used in therapeutic monitoring, in patient follow up procedures and in assessing tumour recurrence or metastasis likelihood. Much before the availability of the clinical or radiological evidence of cancer recurrence, it may be biochemically identified by increasing predictive biomarker levels. Continued follow-up during and after therapy for cancer patients may reflect their condition if biomarker levels have not been elevated or stay at baseline, suggesting effective therapy or remission. The elevation in the level of biomarker above the basal level, on the other hand, suggests disease recurrence. Before all other diagnostic approaches, a predictive biomarker may be a warning indicator of recurrence about as early as 3–12 months. Many biomarkers, such as CEA in CRC cancers, CA125 in ovarian cancers, or PSA in prostatic cancer, may be used to control treatment or diagnose recurrence or metastasis (Kretschmer and Tilki, 2017). As a screening marker for pancreatic cancer, CA19-9 was approved by the FDA in 2002 (Luo *et al.*, 2020).

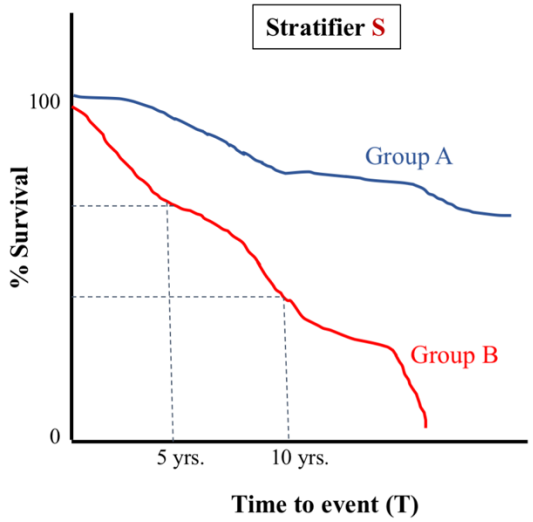
2.4 '*omics*'-based biomarkers and genetic tests

The emerging area of precision medicine in cancer relies on the information provided by a recent field known as *'omics'* which involves several related areas such as genomics, proteomics, transcriptomics, metabolomics etc (Ielapi *et al.*, 2020). The contribution of genomics-which studies the whole genome- and proteomics-which studies the protein repertoire-to precision medicine has gained the greatest share of coverage since the success of the Human Genome Project. Omics tools are high-throughput techniques that generate vast quantities of data about molecules of interest. Examples include next-generation sequencing, for genomics and transcriptomics research, and mass spectrometry used in proteomics. The omics technologies have contributed significantly in the identification of relevant biomarkers in cancer (Olivier *et al.*, 2019). The data retrieved from omics techniques are analyzed to determine biomarker role in specific cancer occurrence, or in cancer prognosis, or in assessing the response to a particular therapeutic intervention. The ability to generate a thorough disease characterization makes it easier to stratify patients into well-defined personalised management and treatment classes, which is the

cornerstone of precision medicine. Various genomics based biomarkers have shown excellent performance in different cancer cohort studies (Quezada *et al.*, 2017). Some of these are commercially available as 'genetic tests' and are increasingly being used at the clinical level. Few prominent examples include OncotypeDX (colon cancer), Prolaris (prostate cancer), Melagenix (melanoma), Mammaprint (breast cancer), SPOT-Light HER2 CISH (breast cancer), ImmunoCyt (bladder cancer), MESOMARK (mesothelioma), OvaSure (ovarian cancer), HybriTech (prostate cancer) etc (Ebell, 2019; Brandao *et al.*, 2019; Mian *et al.*, 1999; Beyer *et al.*, 2007; Kretschmer and Tilki, 2017).

2.5 Cancer management through survival curves

When an individual is diagnosed with cancer, he/she is often presented with a survival estimate depending on the stage of the disease. This is typically assessed by means of a survival plot, wherein in the outcome of a patient population is shown in terms of curves (Rich *et al.*, 2010). The likelihood of survival in 5- or 10-years is the most common retort from a survival plot. A similar



S	A	B
Cancer type	Breast Cancer	Lung Cancer
Age with a specific cancer	Age<60	Age>60
Prognostic biomarker	Good prognosis	Poor prognosis
Predictive biomarker	Therapy responders	Therapy non-responders
Diagnostic biomarker	Without cancer	With cancer
Therapy choice	Therapy T1	Therapy T2

Figure 2.3 Survival curves for patient management in cancer

graph is also utilized to discuss the efficacy of a treatment. It is therefore important to understand how the survival curves are obtained as well as represented. A survival curve is a plot of the fraction of patients as a function of time. Thus, the vertical axis represents the survival chance and the horizontal axis is the time to a certain event such as time to death (overall survival) or time to disease relapse (disease free survival). **Figure 2.3** illustrates a typical survival plot wherein two

population groups A and B are segregated based on the stratifying condition, S. Survival plots are heavily used for assessment of biomarkers and risk factors in a specific cancer population. However, they can also be used amongst groups with different pathological conditions such as two different cancer types.

2.6 Apoptotic pathways

A comprehensive knowledge of the mechanism of Apoptosis is vital to understand the pathogenesis of cancer and other conditions. This understanding, subsequently helps in the identification of key molecules, thereby motivating development of novel biomarkers and/or drugs that target these molecules. As discussed in the earlier chapter, the most central players in apoptotic process are caspases (Li and Yuan, 2008). Caspases by their function as both initiators and executioners drive the cell death events, in response to certain extra-cellular or intra-cellular stimuli. There are two major pathways through which the initiation of caspases takes place (a) the intrinsic or mitochondrial pathway and (b) the extrinsic or death receptor pathway. Both of these pathways converge to a conjoint pathway known as the execution pathway that ultimately activates the executioner caspases (caspase 3/7) and lead the cell to its demise. Details regarding both these pathways is given in the following sections.

2.6.1 The intrinsic or Mitochondrial apoptotic pathway

As the name implies, this pathway is activated due to various intra-cellular stresses such as oxidative stress, radiation, DNA damage etc. The quintessential event for intrinsic apoptosis to take place is the loss of mitochondrial membrane's integrity also known as MOMP (mitochondrial outer membrane permeabilization) (Kim, 2005). MOMP leads to the release of mitochondrial cytochrome-c into the cytosol which binds with APAF1 to form a ring-shaped complex known as Apoptosome (Yuan and Akey, 2013). Apoptosome provides a platform for pro-caspase 9 to caspase 9 conversion, which subsequently activates caspase 3/7 and steers the cell to its fate (Jin and El-Deiry, 2005). The regulation of this process is managed by Bcl2 family of proteins which are divided into two classes: pro-apoptotic proteins and anti-apoptotic proteins (Reed *et al.*, 1996). Pro-apoptotic proteins such as Bax and Bak,,when activated, locate themselves to mitochondrial membrane and cause mitochondrial pore formation. The anti-apoptotic proteins such as Bcl2, Bcl-XL and Mcl1, on the other hand, inhibit pro-apoptotic proteins by binding to them (Ghobrial *et*

al., 2005). There is a specific category of pro-apoptotic proteins known as BH3 only proteins such as Bid, Bim, Puma, Noxa etc. which sense the death stimuli and activate Bax/Bak. Yet another class of proteins exist, known as inhibitors of apoptosis (IAPs) which inhibit the activity of executioner caspases and prevent cell death (Elmore, 2007). A common example is XIAP. However, mitochondrial proteins such as SMAC/DIABLO are known to neutralize IAPs. An illustration of this process is shown in **Figure 2.4.** The aim of any cancer therapy that targets the intrinsic pathway is to cause MOMP and induce death of cancer cells. This is often achieved by restoring the dysregulated balance between constitutive proteins. Additionally, the measurement of protein levels of this pathway can also act as a potential biomarker (Scherr *et al.*, 2016; Charles and Rehm, 2014; Zeestraten *et al.*, 2013).

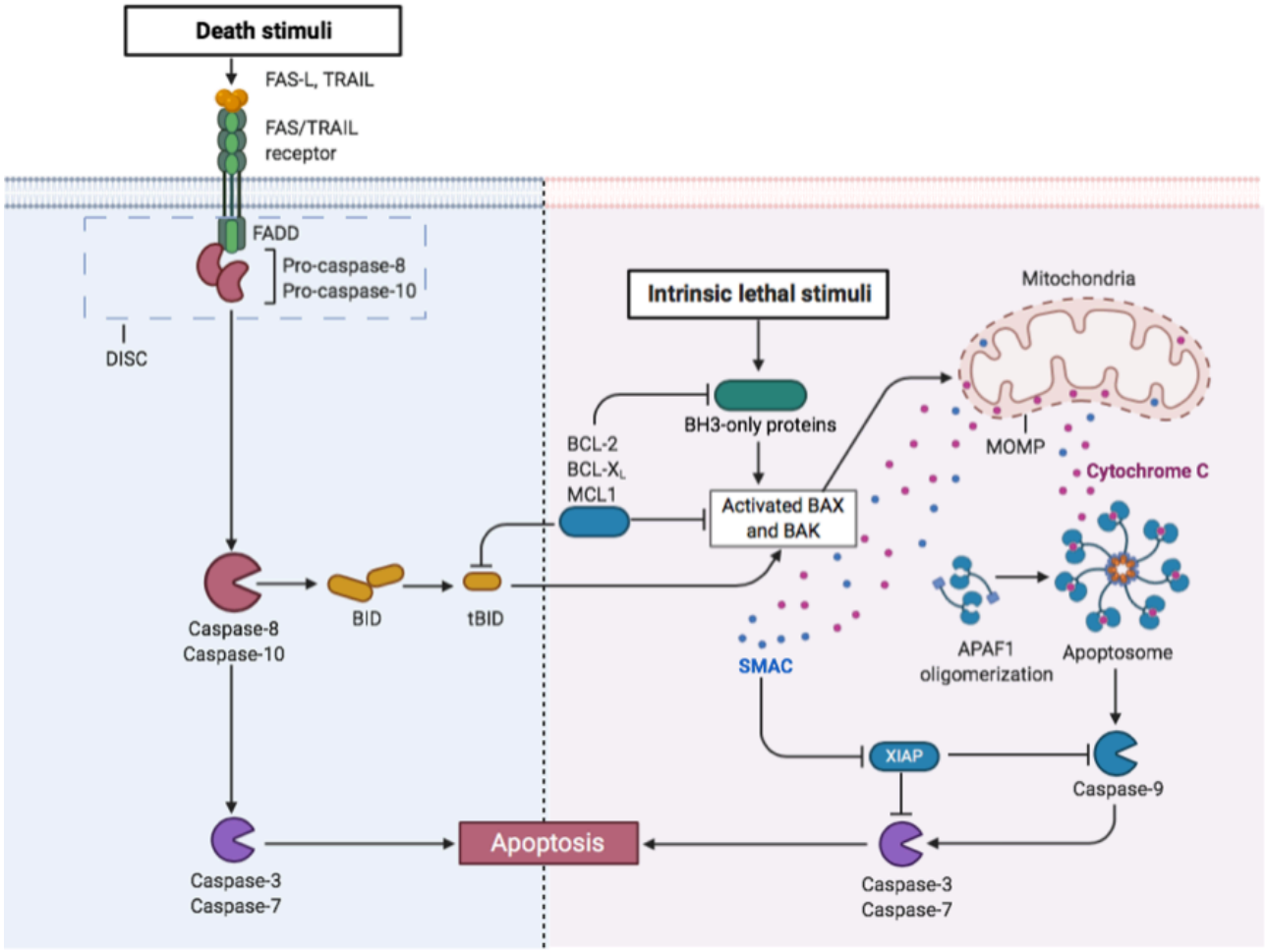


Figure 2.4 Detailed mechanism of pathways involved in the process of apoptosis (source: biorender.com)

2.6.2 The extrinsic pathway

Contrary to the intrinsic pathway, the extrinsic pathway is triggered via extrinsic death stimuli or extracellular death ligands such as FASL (FAS ligand) or TRAIL (TNF-related apoptosis-inducing ligand) in response to extra-cellular environment. These death ligands are recognized by specialized receptors present on the surface of the cell known as death receptors or DRs For example FASL is recognized by FAS (or Apo1) and TRAIL is recognized by TRAIL-R. The process involved in extrinsic pathway of apoptosis is well-established in many of the previous studies (Elmore, 2007; Jin and El-Deiry, 2005; Guicciardi and Gores, 2009). Upon binding to DRs, a DISC or death inducing signaling complex is formed. DISC is a complex which involves proteins such as FADD (DD-containing Fas-associated death domain), procaspases 8/10 and cFLIPS (cellular FLICE inhibitory proteins) (Bredesen *et al.*, 2006). The procspases 8,10 are cleaved to their activated caspase forms leaving their pro-domains on the DISC. The activated caspases thereafter trigger the executionary caspases 3,7 leading to apoptosis. Caspase 8 is also responsible in truncation of Bid to its activated form tBID which then activates the pro-apoptotic molecules Bak/Bax thereby leading to MOMP. **Figure 2.4** clearly represents this process.

2.7 Apoptosis related molecules as cancer biomarkers

Cancer is a highly complex cluster of ailments that represent fundamental anomalies that alternate with natural cell activity including abnormal cell growth and expansion. The biological mechanisms for the development of cancer are generally classified into six processes: proliferative signalling, preventing growth suppression, resistance to apoptosis or cell death, allowing replicative immortality, causing angiogenesis, and finally initiating invasion and metastasis (Hanahan and Weinberg, 2000). The identification and development of novel cancer biomarkers and their increasing therapeutic efficacy in cancer patients can be attributed to a detailed understanding of the altered molecular pathways and cellular processes driving carcinogenesis. Amongst these mechanisms, apoptosis is the most widely studied, which has led to the identification of several key molecules. These molecules have been identified as biomarkers in various cancers. We use the database CIViCmine (Lever *et al.*, 2019) to mine the information

regarding the role of different apoptosis related genes in biomarker development. **Figure 2.5a** shows the role of some major genes as prognostic, predictive and diagnostic biomarkers in various

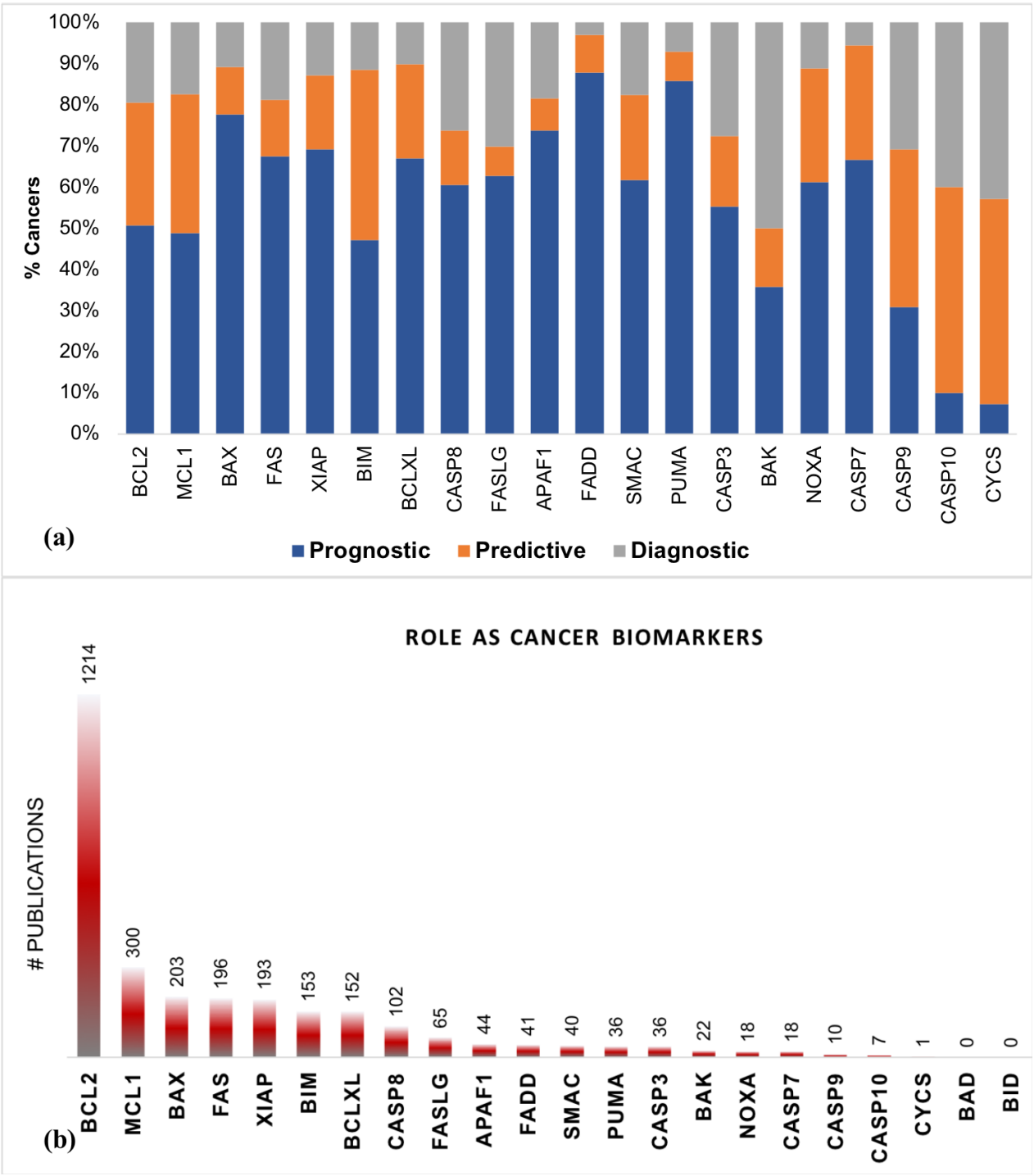


Figure 2.5 The role of apoptotic genes as cancer biomarkers. (a) Figure showing the distribution of roles as prognostic, predictive and diagnostic biomarker across various cancers. (b) The number of publications reporting the biomarker role of the specific gene. (data source: CIViCmine database)

published studies across different cancers. **Figure 2.5b** reports the number of published studies which report these findings. From these results, it is quite clear that the role of Bcl2 as biomarker has been reported the highest number of times with half of the studies mentioning it as a prognostic biomarker. Amongst the pro-apoptotic genes, Bax is seen to be the most reported molecule with majority of the studies mentioning it as prognostic biomarker. Several researchers have also claimed that Bcl2/Bax ratio is an effective indictor of cancer prognosis (Vucicevic *et al.*, 2016; Csuka *et al.*, 1997). Apart from this, a number of studies have also looked at the protein expression profile of apoptosis related molecules and associated their levels with cancer risk. Some recent examples include association of Caspases 3/6, XIAP and APAF1 with patient survival in Melanoma (Charles and Rehm, 2014); the expression of Bcl2, Fas/FasL and TRAIL as prognostic biomarkers in colorectal cancer (Zeestraten *et al.*, 2013); the expression of BIK as biomarker for tumor recurrence in gastric cancer (Pandya *et al.*, 2020); MCL1 expression in lung cancer (Nakano *et al.*, 2020) and other molecules of signalling pathway (Bai *et al.*, 2011; Ding *et al.*, 2020; Zeng *et al.*, 2019; Ma *et al.*, 2019).

2.8 Apoptosis as target in cancer therapy

Each impairment or anomaly along the apoptotic mechanisms can also be an important focus of cancer therapy. Drugs or therapeutic methods that can revert the pathways to normality of apoptotic signalling have the ability to kill cancer cells, which depend on these defects to remain alive. Several recent and substantial results have opened new doors to possible new types of anti-cancer therapies. The use of chemotherapeutic drugs to block the Bcl2 family of anti-apoptotic proteins and the silencing of upregulated proteins or genes involved are several possible treatment methods. The first agent targeting Bcl2 to enter clinical trials was Oblimersen sodium. Examples of drugs that affects expression of Bcl2 family of proteins include ABT-737, ABT-263 and GX15-070. The BH3 mimetics, so called because they imitate the binding of these proteins to the Bcl2 protein hydrophobic groove, are another class of drugs. In animal models with a high percentage of cures, these mimetics have been shown to induce regression of existing tumours. For example ABT 737 has been shown to bind with Bcl2, BclXL and BclW and inhibit their function. It has also been stated that other BH3 mimetics, such as ATF4, ATF3 and NOXA, bind to and inhibit Mcl-1. Apart from this, some experiments have shown that an improvement in apoptosis could be accomplished by silencing genes coding for the Bcl-2 family of anti-apoptotic proteins. Silencing

Bmi-1 in MCF breast cancer, for example, was shown to make the cancer cells more vulnerable to doxorubicin. Further, the most potent apoptosis inhibitor of all IAPs has been reported to be XIAP. Antisense methods and short interfering RNA (siRNA) molecules are some of the experimental therapies targeting XIAP. Using the antisense method, XIAP inhibition is documented to result in increased in vitro radiotherapy regulation of tumours. XIAP antisense oligonucleotides have been shown to

show elevated chemotherapeutic function in lung cancer cells if used simultaneously with anticancer drugs. Lastly, several therapeutic agents have been developed to activate caspases

Table 2.2 The list of drug molecules that target apoptotic pathway.

Drug	Alias	Target	Used in combination with	Benefactor	Cancer/condition
ABT263	-	Bcl2 family	erlotinib/irinotecan	Abbott	Solid cancers
ABT263	-	Bcl2 family	docetaxel	Abbott	Solid cancers
ABT263	-	Bcl2 family	paclitaxel	Abbott	Chronic lymphocytic leukaemia
ABT263	-	Bcl2 family	-	Genetetch	Chronic lymphocytic leukaemia
AT101	Gossypol	Bcl2 family	-	Roswell park cancer institute	Chronic lymphocytic leukaemia, Chronic B-cell leukaemia
AT406	-	IAPs	-	Ascenta	Solid cancers, Lymphoma
AT406	-	IAPs	-	Ascenta	Acute myelogenous leukaemia
ENZ3042	-	IAPs	-	Therapeutic advances in childhood leukaemia consortium	Acute, childhood, T cell lymphoblastic leukaemia
GX15070MS	Obotoclax	Bcl2 family	-	Children's oncology group	Leukaemia, Lymphoma
GX15070MS	Obotoclax	Bcl2 family	-	Arthur G James cancer hospital and Richard J Solove research institute	Lymphoma
HGS1029	-	IAPs	-	Human Genome Sciences	Solid cancers
HGS1029	-	IAPs	-	Human Genome Sciences	Solid cancers
LCL161	-	IAPs	-	Novartis	Solid cancers
RO5458640	-	TWEAK ligand	-	Hoffmann-La Roche	Solid cancers

synthetically. For example Apoptin, originally derived from the chicken anaemia virus, is a caspase-inducing agent. In some trials, caspase-based gene therapy has been attempted in addition to caspase-based drug therapy. It was observed that the effects of this therapy caused significant apoptosis and reduced the volume of the tumour. **Table 2.2** lists some molecules that are undergoing/completed clinical trials and target apoptosis.

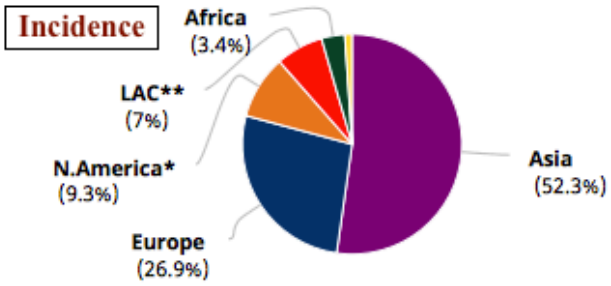
3

Risk Prediction using Protein Expression Profile

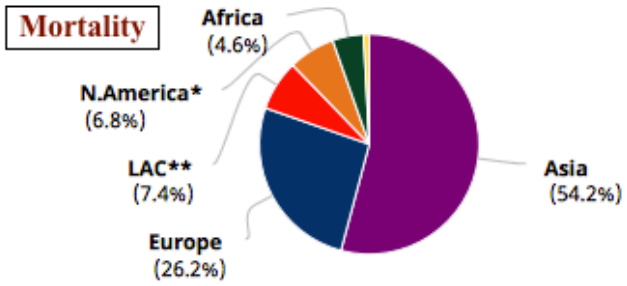
Colorectal Cancer

3.1 Introduction

Large bowel cancer or Colorectal cancer (colon and rectum) is one of the most lethal cancers with the second largest death rate amongst all cancers in the west. According to the latest colorectal cancer statistics provided by the American Cancer Society (Siegel *et al.*, 2020), in US alone, around 147,950 incidences and 53,200 deaths are estimated for the year 2020 which includes 3,640 deaths in people with age less than 50 years. It is also observed that while incidence and mortality rates have shown a decline in the age group of more than 50 years, an increase in both these rates has been seen for individuals with age less than 50 years. Globally, around 10% of all the deaths due to various cancers have been attributed to colorectal cancer, for the year 2020 (Global Cancer Observatory). Also, amongst the number of deaths due to colorectal cancers, Asian countries account for the maximum number of deaths. The number of incidences also follow a similar pattern (**Figure 3.1**). The GCO database from WHO provides updated fact-sheets about information regarding country-wise incidences and mortality rates based on sexes and age groups. The cause for this geographic disparity has generally been associated with diverse dietary habits across the world as well as distinct environmental exposures. Other lifestyle related factors (such as physical



	Population	Number
	Asia	1 009 400
	Europe	519 820
	*Northern America	180 575
	**Latin America and the Caribbean	134 943
	Africa	66 198
	Oceania	20 654
	Total	1 931 590



	Population	Number
	Asia	506 449
	Europe	244 824
	**Latin America and the Caribbean	69 435
	*Northern America	63 987
	Africa	42 875
	Oceania	7 603
	Total	935 173

Figure 3.1 Global incidence and mortality rates of colo-rectal cancer (Data source: GCO,WHO)

activity, obesity, alcohol intake and smoking) have also been reported to play an important role in CRC development and progression.

Once the individuals encounter some specific symptoms such as rectal bleeding, vomiting, obstruction, pain etc. related to different colon or rectum sites (**Figure 3.2**); they are advised by the doctors to undergo certain screening procedures for CRC detection. Detecting CRC at its earliest stage provides the greatest chance for a cure. The most common screening procedures used

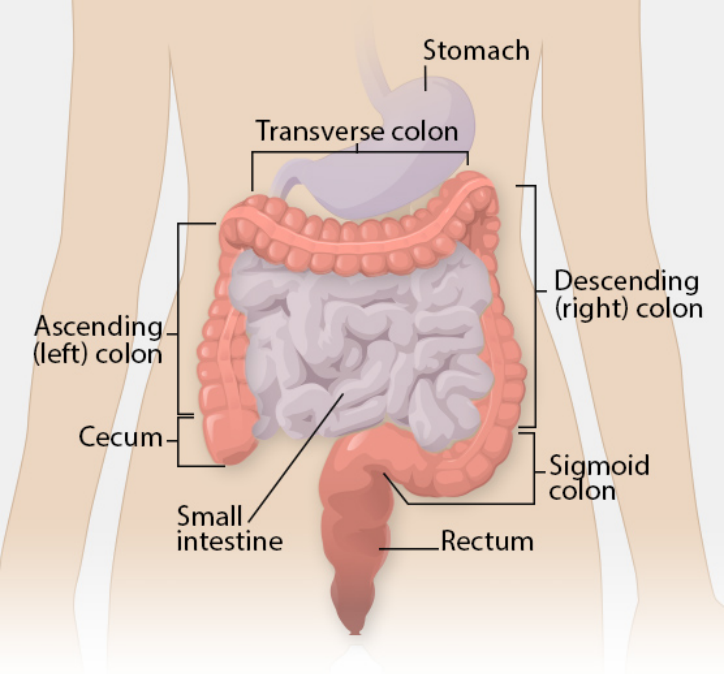


Figure 3.2 The parts of colon and rectum

(Image from https://www.cdc.gov/cancer/colorectal/basic_info/what-is-colorectal-cancer.htm, CC BY-SA 4.0 via Wikimedia Commons)

currently include faecal occult blood testing, digital rectal examination and sigmoidoscopy. Screening has been shown to reduce the risk of mortality due to CRC, however, clinicians generally recommend people with an average risk of CRC begin screening around age 50. Screening is followed by staging for which Duke's staging method is most widely used and information regarding other prognostic factors such as degree of penetration of the primary tumour, lymph node involvement, resection margins, vascular or lymphatic invasion and large-bowel obstruction is also considered. Once the cancer extent is established, patients undergo various treatment procedures depending on their situation and health. For early stage CRC patients surgical resection procedures such as polypectomy, laparoscopy, endoscopic mucosal resection

and partial colectomy are employed. For advanced stages, surgery is either followed by chemotherapy or vice-versa.

Chemotherapeutics are the class of drugs which target specific cellular targets and destroy cancer cells. Mainstay chemotherapeutics for CRC treatment include 5-FU, oxaliplatin and irinotecan. While, the ongoing advancement in the field of genomics is leading the development of novel chemotherapeutics, the success rate of these chemotherapeutics is not at par. The failure of a therapy adds to both health (due to toxic side effects of chemotherapy) and financial burden on the lives of the patients. As a solution this problem, modern clinicians usually evaluate the therapy's success/failure rate based on certain biomarker levels in the patient's body. Several of these predictive biomarkers have been previously established (Lee and Chan, 2011; Koncina *et al.*, 2020) for CRC. Predictive biomarkers measure the likelihood of response or lack of response of a particular therapy, and allow identification of patients most likely to benefit from a given treatment, thus sparing other patients from toxicities of ineffective therapies. These biomarkers are majorly molecules associated with biological conditions or processes that are inconsistent amongst cancer and non-cancer population. One such condition is dysregulated cell death or apoptosis process in cancer. The insights achieved from understanding the apoptosis process in cancer has shed light on the variation in the expression of Bcl2 family proteins, its role in tumorigenesis and prognosis (Yang *et al.*, 2009; Stoian *et al.*, 2014; Liao *et al.*, 2018; Yi *et al.*, 2016; Li *et al.*, 1998). A recent mathematical study, involving the proteins of mitochondrial type 2 pathway, has also proposed a predictive biomarker for therapy response in CRC (Lindner *et al.*, 2013; Andreas U. Lindner *et al.*, 2017).

In regard to this, the current study utilized a dataset containing Stage III CRC cohort of patients which have undergone Xelox and Folflox chemotherapy regimens. Both Xelox and Folflox are oxaliplatin based drugs used for advanced stage patients. Oxaliplatin is known to cause DNA damage in colon/rectal cancer cells, thereby inducing Bax translocation to mitochondrial membrane and ultimately resulting in cell demise. However, the crucial balance between the pro-apoptotic and anti-apoptotic proteins can be a decisive factor for the efficacy of these chemotherapy regimens. The failure of these regimens can lead to a huge burden on patients in termns of toxic side effects and a significant loss of resources. The prediction of therapy outcome can be a huge development in the CRC patient management. Since, the therapies are mainly based

on protein function, the prediction method needs to be developed on protein expression data. This method should further encompass the intricate proapoptotic-anitapoptotic balance. In this study, we gauged the predictive potential of expression of proteins from the Bcl2 family in stratifying patients into high risk (non-responder) and low risk (responder) groups. By means of various statistical and machine learning models, we established a protein signature which can be used to predict the response of Xelox/Folflox chemotherapy and provided a comprehensive comparison with clinical factors and another popular biomarker. We developed a web-based tool, to provide service to the community, which can be utilized by clinicians for classifying patients into risk groups. Further, by utilizing an external web-based tool, we show that the Bcl2 protein expression data can also stratify stage III colon and rectal patients into high/low risk groups beforehand.

3.2. Materials and methods

3.2.1 Dataset and pre-processing

The 'CRC stage III cohort' dataset used for this analysis was derived from (Andreas U. Lindner *et al.*, 2017). The dataset was retrieved by permission from the authors on 24th Sept 2018. It includes information from primary tumour samples of Formalin-fixed paraffin-embedded (FFPE) from 134 subjects treated with FOLFOX and XELOX therapy regimens. In particular, it comprises of Bcl2 family protein levels in Nano-Molar (nM) retrieved by reverse-phase protein array (RPPA). Additionally, the dataset also provides full clinical details such as overall survival (OS) time, censoring information, lympho-vascular invasion, M staging; gathered from patients' medical surveillance. The data was normalized prior to further analysis. This dataset (n=134) was further used to compare the model developed in this study with a previous model DR_MOMP as applied to the same dataset

3.2.2 Model development and conceptualization of 'Risk Score'

Multiple linear regression (MLR) models using Python's sklearn package (v0.20.3) have been applied to fit protein expression levels with the OS time. Ordinary least squares, Lasso regression, Ridge regression, Lasso-Lars regression, Bayesian ridge regression and Elastic-net regression

models are the approaches used to approximate the regression coefficients. The model training and test evaluations were executed out by means of a five-fold cross-validation procedure. The incorporation of all "five" predicted test datasets (predicted OS) was utilized to categorise the actual survival time (OS) at mean/median cut-offs using Cox survival analysis. Coefficient optimization and regularisation have been accomplished using built-in approaches such as RidgeCV, LassoCV, LassoLarsCV, etc.

We also implemented a parameter optimization technique, wherein, the coefficients (w) of the sum β defined (for a given sample) as

$$\beta = w_{Bak}[Bak]+w_{Bax}[Bax]+w_{Bcl2}[Bcl2]+w_{BclXL}[BclXL]+w_{Mcl1}[Mcl1]$$

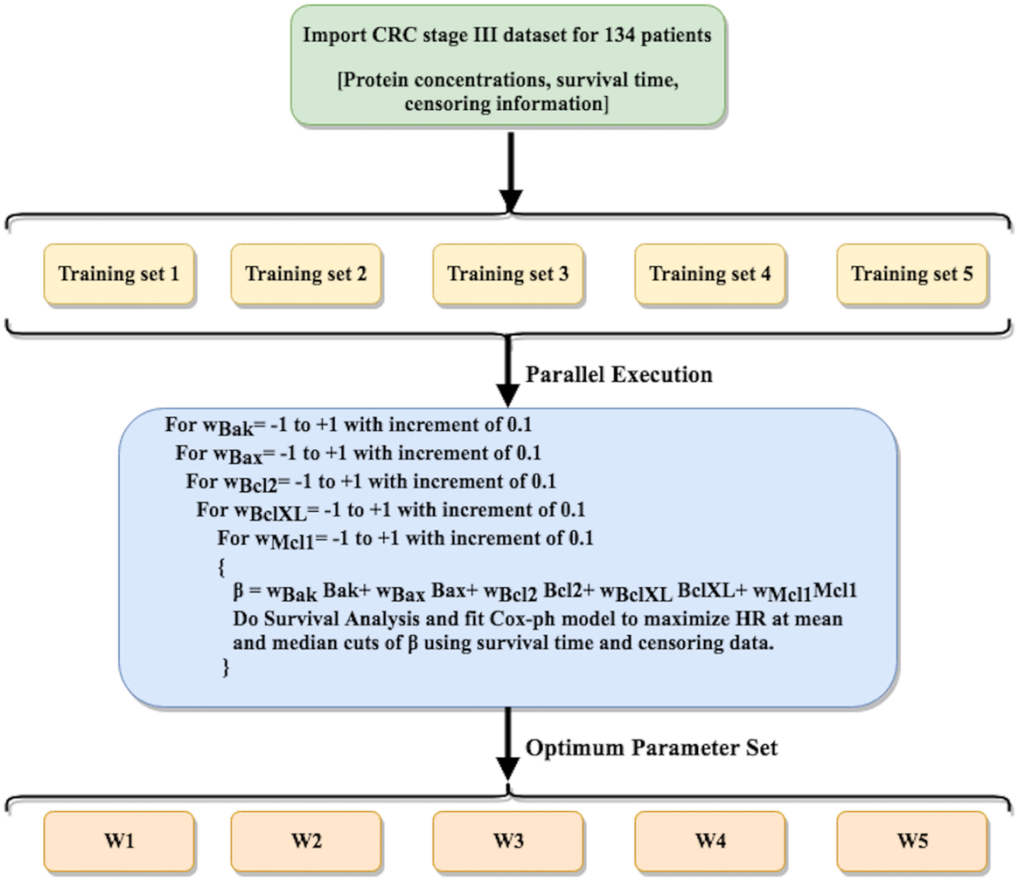


Figure 3.3 Pseudocode for parameter optimization technique (doi: 10.1371/journal.pone.0217527)

were optimized using five training sets derived from the complete dataset. For each training set i, a β^i with optimized coefficient set w^i, was obtained which maximized the objective function

'Hazard Ratio' at mean and median cut-offs. The pseudocode for the algorithm is shown in **Figure 3.3**. Subsequently, β* was constructed where each coefficient is taken to be the mean of five coefficients obtained from the training sets earlier, For example

$$w_{Bak}^{*}=(w_{Bak}^{1}+ w_{Bak}^{2}+ w_{Bak}^{3}+ w_{Bak}^{4}+ w_{Bak}^{5})/5$$

The standardized version of β* was termed as 'Risk Score'.

3.2.3 Evaluation metrics

Hazard ratios (HR) and Confidence intervals (CI) were calculated to estimate the probability of mortality linked with high-risk/low-risk classes stratified with the univariate unadjusted Cox-PH models on the basis of mean/median values of different variables. In order to better evaluate various covariates, multivariate Cox-PH models were used to determine the relative death risks related to various variables. In order to assess the survival curves of high- and low-risk factions, Kaplan-Meier (KM) plots were employed. Survival tests were conducted using the 'survival' library in R on these datasets. Using log-rank tests, statistical significance was calculated between the survival curves. In order to measure the value of the explanatory variables used for HR estimates, Wald tests were conducted.

3.3 Results

3.3.1 BclXL protein expression as biomarker

We performed a Cox-PH univariate survival analysis using the numerical variables provided in the dataset i.e. protein levels and patient age. Based on these multiple single variables at the median cutoff, we segregated high and low risk patients. According to this analysis, the findings in **Table 3.1** indicate the HR, CI and p values. On the basis of each protein's concentration, we calculated hazard ratios and CIs to analyse if either of them would serve as a predictive marker that distinguishes responsive or low risk patients with non-responsive or high risk patients. HR spanned from 1.3 (age) to 20.877 (BclXL), as seen in **Table 3.1**. On the basis of both mean (HR = 7.19, p-value = 0.0004) and median (HR = 20.81, p = 0.0030) cut-offs, BclXL was capable of separating high and low risk CRC patients, thereby reaching optimum distinction.

Table 3.1 The performance of univariate survival models developed on different variables and their combination; BclXL showed the highest performance.

CRC Stage III (n=134)	Median Cutoff	
Variable	**HR (%95CI)**	**p-value**
Age	1.3 (0.54-3.14)	0.55
Bax	1.34 (0.55-3.23)	0.52
Bak	2.79 (1.07-7.3)	0.04
Bcl2	1.25 (0.51-3.02)	0.62
BclXL	20.81 (2.7-155.5)	0.003
Mcl1	1.64 (0.67-4.03)	0.27
Bcl2+BclXL+Mcl1+Bax+Bak	6.37 (1.86-21.73)	0.003
Bcl2+BclXL+Mcl1-Bax-Bak	2.49 (0.95-6.47)	0.06

*CI: Confidence Interval, HR:Hazard ratio, samples>median (variable) were taken as high-risk group

3.3.2 Multiple linear regression models for risk assessment

We measured variations in mean concentrations of all proteins amongst patients who survived the trial and patients who succumbed to death or those whose cancer relapsed in order to claim BclXL as an exclusive predictive biomarker. On each of these proteins, a t-test was conducted, and it was found that Bak ($p = 0.0042$), Bax ($p = 0.0094$), BclXL ($p = 3.5e\text{-}05$) and Mcl1 ($p = 0.02$) levels varied significantly between the two classes.

Table 3.2 The peformance of prognostic models developed using regression based techniques on multiple variables.

Model	Mean Cutoff		Median Cutoff	
Name	**HR**	**p-value**	**HR**	**p-value**
LR	3.19	0.0132	3.27	0.0219
Ridge	3.34	0.0101	3.27	0.0219
Lasso	1.79	0.196	2.09	0.1170
LassoLars	2.44	0.0472	6.34	0.0032
Elastic-net	1.79	0.1960	2.15	0.1030
Bayesian ridge	2.08	0.1010	2.64	0.0469

*HR: Hazard Ratio, LR: Linear Regression, patients with <mean (predicted OS) or <median (predicted OS) were taken as high risk group.

In conjunction with BclXL, this finding demonstrated the importance of other proteins and dismissed the idea of using BclXL as the exclusive biomarker. As a result, we used the total protein concentration (sum) and difference between anti-apoptotic protein levels and pro-apoptotic protein levels to stratify risk groups. The results depicted in **Table 3.1** show that total concentration levels were able to classify the two classes with a maximum HR = 6.37, p-value = 0.0030 at the median cutoff. This motivated us to use multiple linear regression models with protein levels as independent variables and OS as target or dependent variable. It was found that the model based on LassoLars worked better than other models and obtained a maximal HR value of 6.34 with p-value = 0.0032 at the median cutoff using the the predicted OS (**Table 3.2**). Although this approach utilized multiple protein data and offered predicted OS as a predictive biomarker that performs better than many previously developed markers, it still underperformed in contrast to BclXL levels.

Table 3.3 Hazard Ratio (HR) of prognostic models developed using parameter-optimization technique. Risk Score (RS) was computed using a simple linear function by optimizing weights.

Case	Mean		Median	
$w=(w_{Bak}, w_{Bax}, w_{Bcl2}, w_{BclXL}, w_{Mcl1})$	HR	p-value	HR	p-value
Set1, w=(0, 0.2, -0.1, -0.8, -0.9)	33.23	0.0006	22.96	0.0023
Set2, w=(0, 0.1, -0.1, -0.9, -0.3)	18.88	8e-05	22.96	0.0023
Set3, w=(0, 0.2, 0, 0.9, 0)	15.94	0.0002	21.54	0.0028
Set4, w=(0, 0.2, -0.2, -0.9, -0.8)	11.26	0.0001	22.41	0.0024
Set5, w=(0.1, 0, -0.1, -0.7, -0.7)	11.03	0.0001	10.35	0.0017
Overall, w=(0.02, 0.14, -0.1, -0.84, -0.54)	38.13	0.0004	22.27	0.0025

*Samples with <mean or <median cutoff were taken to be as high risk group, w is the set of coefficients for different proteins

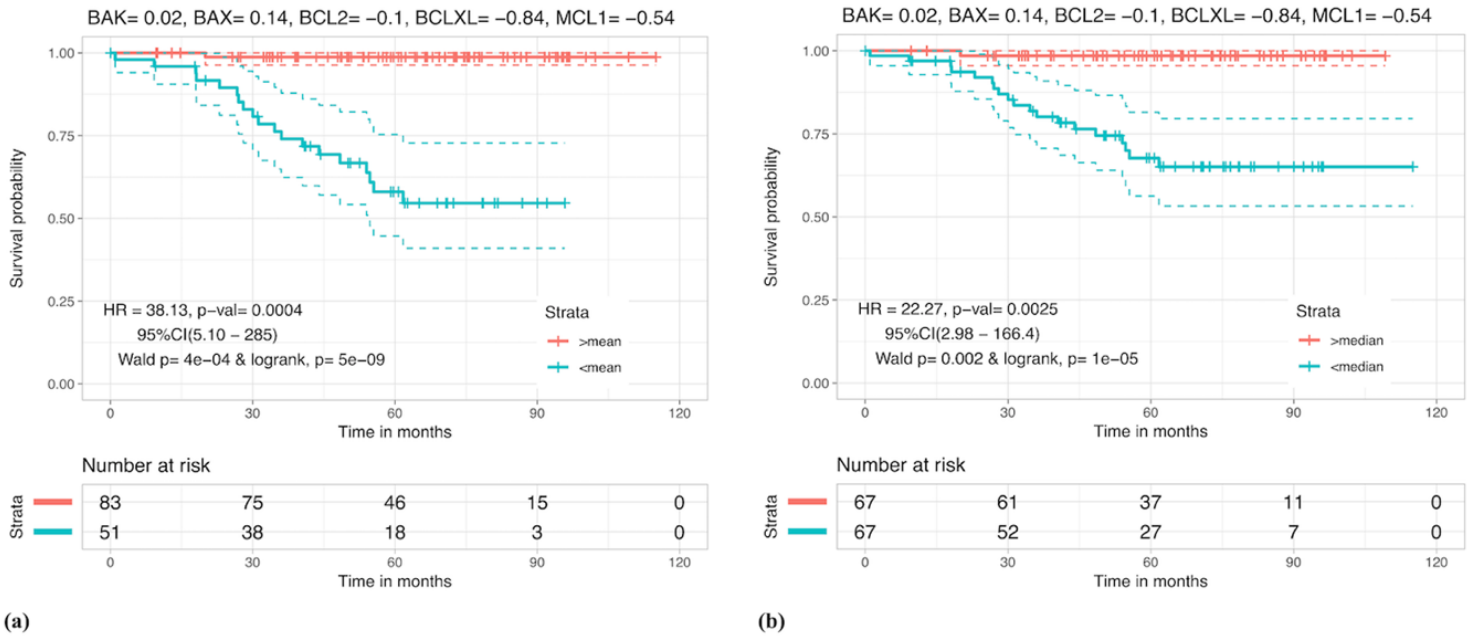


Figure 3.4 Kaplan Meier risk prediction survival curves for CRC patients, based on mean (RS = 0) and median (RS = 0.266) cutoffs. (a) The risk of patients with "RS < 0" was approximately 38 times higher compared to patients with "RS ≥ 0" (HR = 38.13, p = 0.0004). (b) In patients with "RS < 0.266", the risk was nearly 22 times higher than in patients with "RS ≥ 0.266" (HR = 22.27, p = 0.0025). (doi: 10.1371/journal.pone.0217527)

3.3.3 Risk Score (RS) as the most significant biomarker

As outlined earlier, RS was constructed by a parameter optimization technique wherein different training sets were utilized to optimize weights for protein concentrations. The results for different subsets are summarized in **Table 3.3**. Patients with RS < 0 (mean) and RS < 0.266 (median), are found to be at higher risk with HR = 38.13 (p-value = 0.0004) and 22.27 (p-value = 0.0025) respectively, than patients with RS ≥ 0 and RS ≥ 0.266. Kaplan Meier plots for this case are shown in **Figure 3.4**. The number of samples in high/low risk group after the stratification is performed are provided at the bottom of respective KM plots with the title 'Number at risk'. The red line displays the samples in low risk group and blue is representative of samples in high risk group

The weights in Table 3.3 are reflective of the contribution of each of the apoptotic proteins in the sum (β). It was observed that the coefficients obtained for the pro-apoptotic proteins (Bak and Bax) in the linear sum RS were positive, whereas, the coefficients for anti-apoptotic proteins were negative. Further, it was seen that a decrease in RS (due to increase in anti-apoptotic proteins) increases the survival risk (HR>>1). Biologically, this implies that the when the concentration of

anti-apoptotic proteins is higher, the process of apoptosis comes to a halt. This might be a strategy employed by cancer cells to avoid their elimination. Subsequently, the risk of death is increased.

3.3.4 Risk Score vs. Clinical features

A multivariate analysis using cox proportional hazard models, was performed to see the association of other clinico-pathological features present in the dataset with the mortality risk of patients. The findings for mean cutoff are reported in **Figure 3.5**, clearly showing that RS exceeds every other predictor in terms of OS based distinction of patients. RS is shown to be associated with nearly 30 times elevated mortality risk in "high-risk patients" as compared to "low-risk patients" in CRC cohort (HR = 29.44, p-value = 0.001) in the case of mean cuttoff. RS also stratified clinical risk groups as shown in **Figure 3.6**.

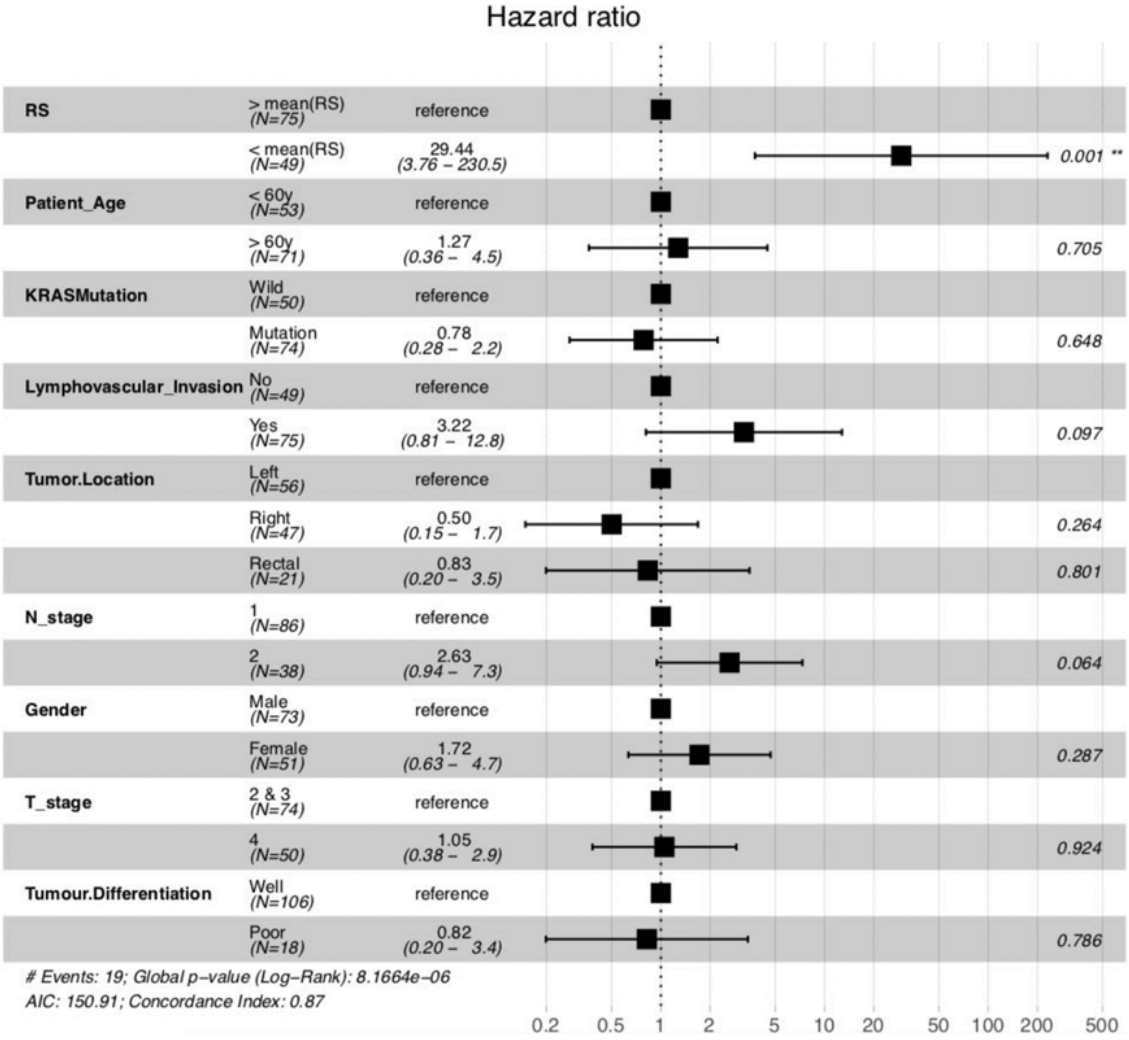


Figure 3.5 RS is revealed as the most significant covariate in the Multivariate survival analysis. (doi: 10.1371/journal.pone.0217527)

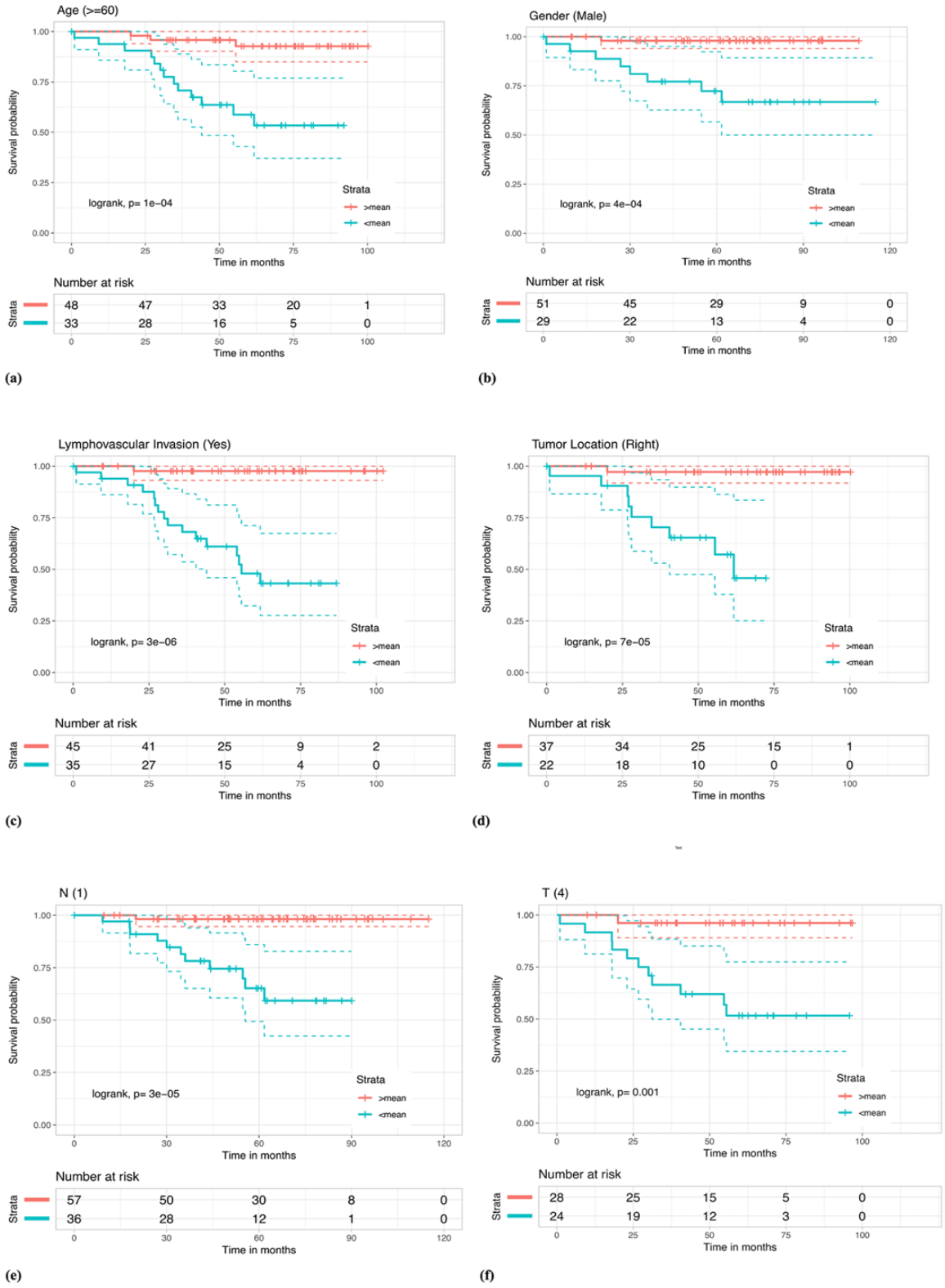
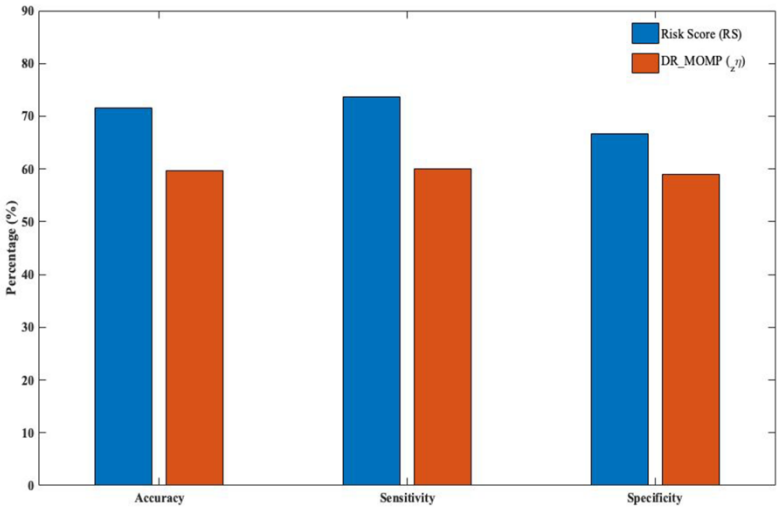


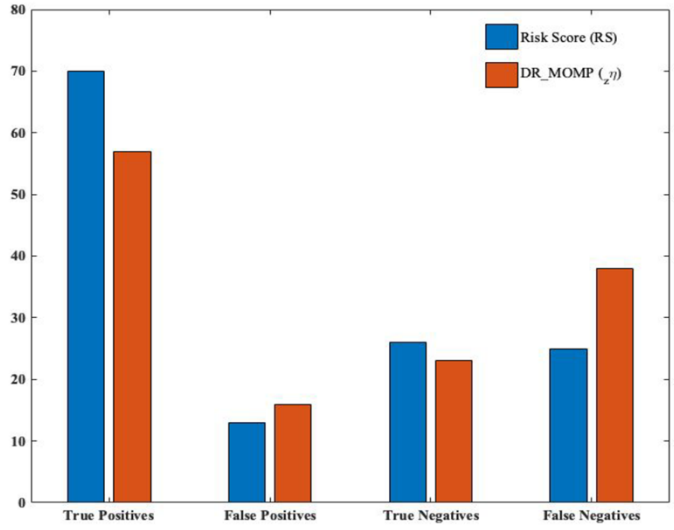
Figure 3.6 KM plots representing the sub-classification of clinical risk groups by RS (mean cutoff) (a) Patients with age>60 (HR=8.04, p=0.0017) (b) Males (HR = 15.91, p= 0.0091) (c) Positive lymphovascular invasion (HR = 24.92, p=0.0018) (d) Righ tumor location (HR = 20.16, p=0.0046) (e) N1 stage patients (HR = 21.11, p=0.0046) and (e) T4 stage patients (HR = 13.65, p=0.0124). (doi: 10.1371/journal.pone.0217527)

3.3.5 Comparison with existing tool

Recently, 134 chemotherapy-treated stage III CRC patients have been graded into risk (responder/non-responder) categories using the DR_MOMP model. The high-risk group identified by DR_MOMP was found to have nearly five times the risk of mortality (HR= 5.2, p-value = 0.02) relative to the low-risk group. (Andreas U Lindner *et al.*, 2017). The CRC stage III cohort dataset contains an additional recurrence information stating that 95 patients were alive during the 5-



(a)



(b)

Figure 3.7 Comparative assessment of RS with DR_MOMP (a) Improvement in sensitivity (73.68%), specificity (66.66%) and accuracy (71.64%) by using RS, as compared to sensitivity (60%), specificity (58.9%) and accuracy (59.7%) of DR_MOMP's zη. (b) Corresponding improvement in prediction of responders (RS>0) and non-responders (RS<=0) with reduced false positives/negatives. (doi: 10.1371/journal.pone.0217527)

-years study period and 39 patients with cases of recurrences/deaths. A comparison between $_z\eta$ of DR_MOMP, and RS was performed on the basis of prediction of recurrence/death vs survival outcomes when concentrations of apoptotic family proteins are known. RS showed a prediction accuracy of 71.64% at mean cutoff, as compared to 59.7% of $_z\eta$. Results are summarized in **Figure 3.7**.

3.3.6 External validation and biological support

To the best of our knowledge, an external dataset with quantified protein expression data as well as survival data for CRC patients was not available. Therefore, to validate our findings we utilized an external web-tool TRGAted (Borcherding *et al.*, 2018), which utilizes Level 4 data from the reverse-phase protein arrays for each cancer type were downloaded from the TCPA Portal (Date Downloaded: 11/10/17) to predict risk groups corresponding to OS in various cancer types. We selected the proteins "BCL2, BCLXL, ,BAK and BAX" for Stage III COAD and READ patients. It is to be noted that MCL1 expression was not available. **Figure 3.8** shows the survival plots for risk stratification based on these proteins. Significant HR was observed in both the cohorts (COAD: 0.291^{-1}=3.43 and READ: 4.24) corroborating our findings. The lower HR as compared to RS on the previous dataset is posiibly due to absence of MCL1 expression. This is evident from the coefficients in RS, which implies that while Bak, Bcl2 and Bax are somewhat less relevant for prognostic studies, BclXL and Mcl1 on the other hand, are the two dominating proteins to look at while stratifying CRC patients. These results also correlate with isolated studies on BclXL and Mcl1 which showed their relevance as prognostic markers in the past (Krajewska *et al.*, 1996; Cho *et al.*, 2017). Several other studies in the past have shed light on the key roles of Bcl2 family proteins in colorectal cancer. In one of these studies, the small molecule drug ABT-737, which inhibits BclXL and Bcl2, was used to culture human CRC tissue ex vivo. The number of apoptotic tumour cells increased considerably after treatment with ABT-737 compared to controls, whereas proliferation levels remained unchanged. The study concluded that Bcl-xL is a driver in colorectal carcinogenesis and cancer development and is a valuable therapeutic target (Scherr *et al.*, 2016). In another study, it was shown that Apigenin which is a natural flavoid, induced the apoptosis of colon cancer cells by inhibiting the phosphorylation of STAT3 and consequently downregulating the anti-apoptotic proteins Bcl-xL and Mcl1 (Maeda *et al.*, 2018). Many other studies such as

(Jokinen and Koivunen, 2015; Tong *et al.*, 2017) also confirm the dominant anti-apoptotic roles of BclXL and Mcl1 in colon cancer.

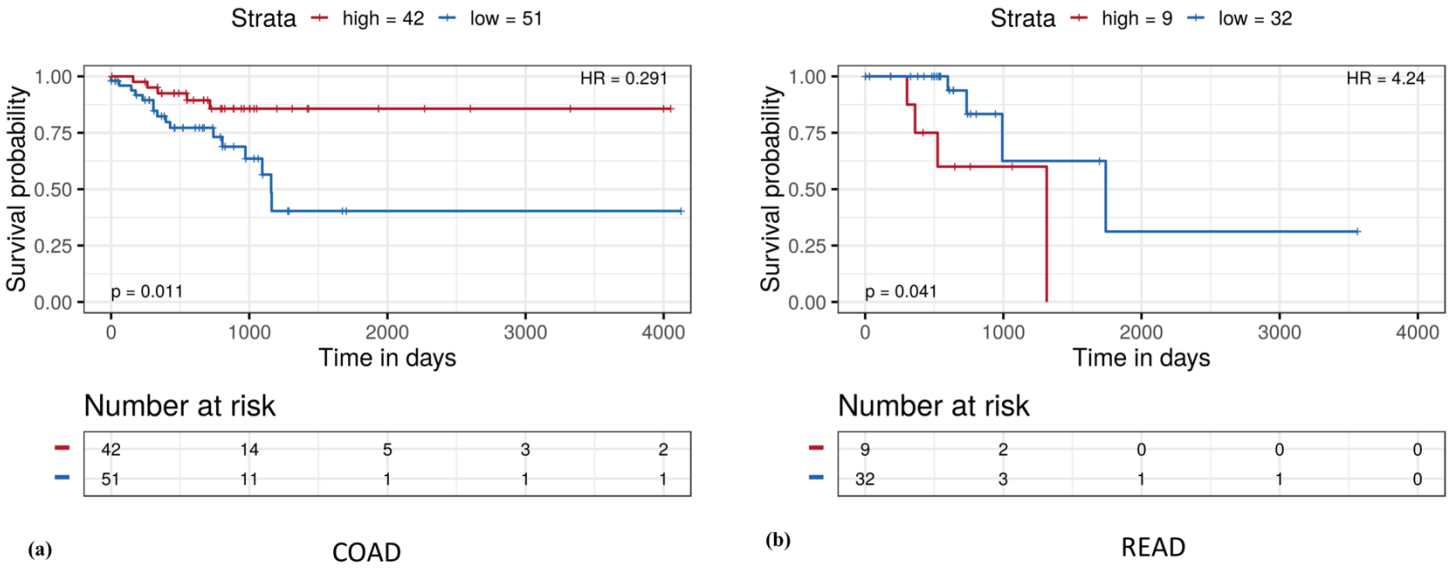


Figure 3.8 Figure shows the result of stratification of Stage III patients by RPPA data of Bcl2, Bax, Bak and BclXL utilizing TRGAted web-tool (a) Risk stratification of COAD patients (b) Risk stratification of READ patients. (source: TRGAted)

3.3.7 Combining RS and patient age enhances stratification

In order to see if certain hybrid combinations could further add significance to the existing protein-based RS, clinical variables were added to RS and allotted optimised weights through an iterative process as before. Various combinations of single features and multi-features were attempted and it was noticed that the combination of β* with age showed the most impactful modification of all single and multiple-feature combinations and used the least number of clinical variables. Adding more attributes to this quantity did not alter the output. This combination was called the Hybrid Risk Score (RS_H). An HR value of 40.11 and a p-value of 0.0003 was obtained for patients with RS_H>median(RS_H). **Figure 3.9** displays the KM plot referring to the stratification of patients by RS_H. This hybrid combination also enhanced the accuracy of the estimation of favourable/unfavourable predictions by 1.5% to 73.13% with sensitivity and specificity values of 75.78% and 66.66%, respectively.

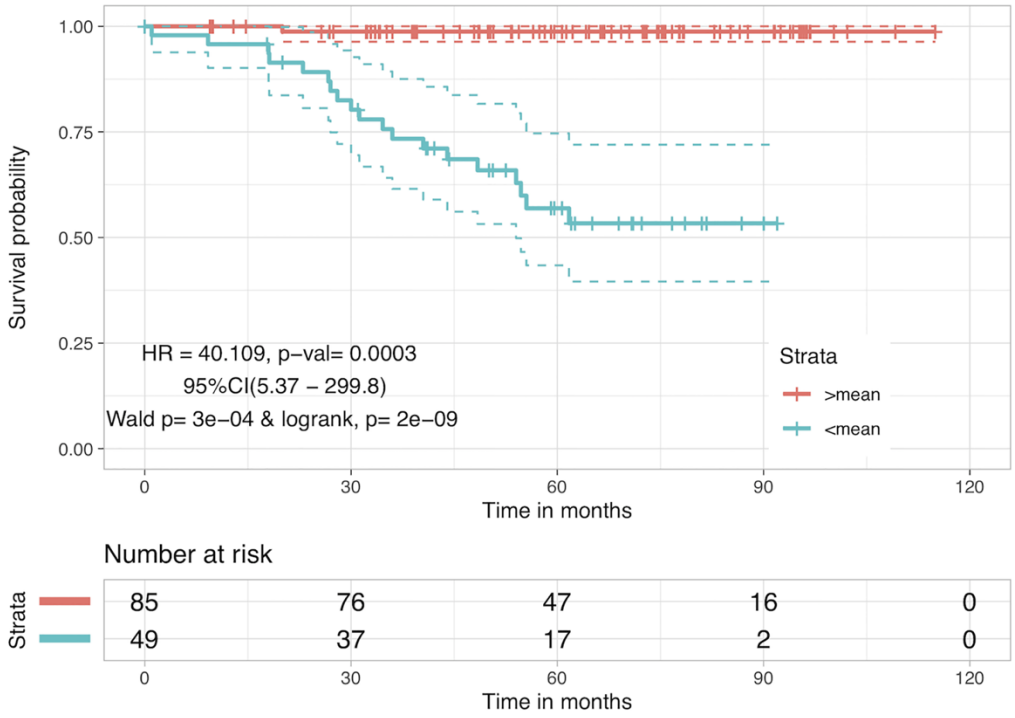


Figure 3.9 Combination of RS with Patient age improves risk stratification. (doi: 10.1371/journal.pone.0217527)

3.4 Web Service and functionality

In order to provide support to the society, we built a web server 'CRCRpred', which is freely accessible at https://webs.iiitd.edu.in/raghava/crcrpred. In order to predict responders (low-risk) and non-responders (high-risk) Stage III patients, given the expression levels of the necessary Bcl2 family proteins, this web server implements the current analysis. **Figure 3.10** displays the basic features, and the two prediction modules with their brief explanations as follows:

3.4.1 Single-protein prediction

Sometimes, the user would not have the concentration of all the necessary Bcl2 family proteins, mostly because the quantification of protein levels is a difficulty in itself. Keeping this in mind, we included this module for the user where, with minimal knowledge of the concentration(s) of one or more proteins, the risk probability can be estimated. The output here is a protein-wise

prediction. The input concentration is supplied to a linear regression model and the risk probability is calculated. This model consists of fitting "bin-wise" mean protein levels with the likelihood of high-risk patients in the bin. High risk and low risk stratification of patients was conducted on the basis of median OS in the CRC cohort.

3.4.2 Multiple-proteins prediction

This module measures a patient's risk score (RS) based on the RS calculation of all five proteins for the patient. Thus the concentration of all five proteins must be known before-hand. The patient is listed in the high/low risk group on the basis of the cutoff, RS = 0. The gap from the cut-off point is presented to the user as a percentage of risk along with the risk score.

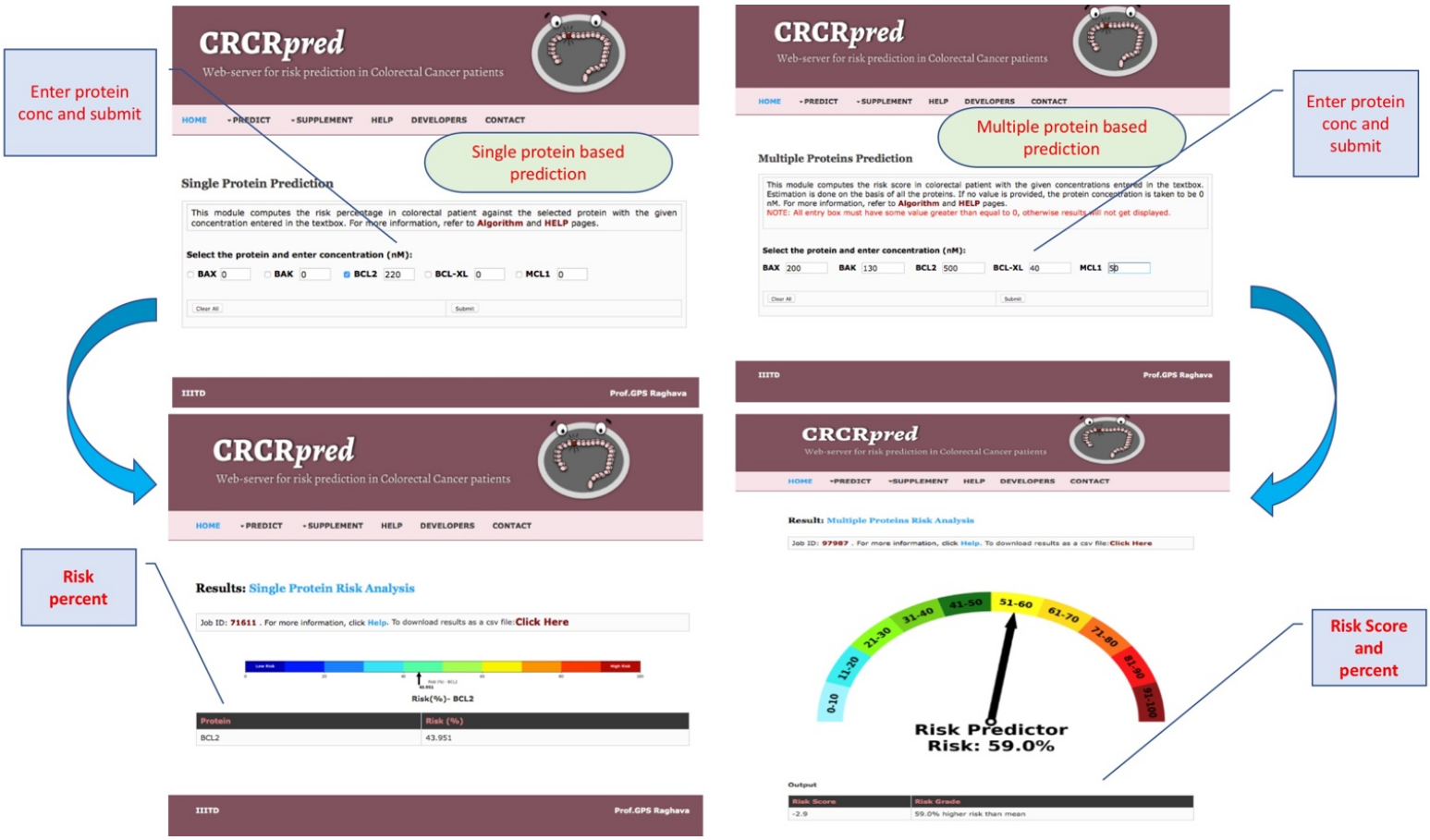


Figure 3.10 Usage of web-service "CRCRpred" for risk estimation in CRC patients by Bcl2 family protein expression data

3.5 Conclusion and Summary

Colorectal cancer is a life-threatening illness with worldwide prevalence that needs improved treatment and patient management techniques. This improvement is only possible if patient-selective therapy or personalized therapy decisions are made. In the past expression profile of apoptotic regulators, specifically of Bcl2 family proteins, has been linked with CRC prognosis and carcinogenesis. Monitoring the protein profile of this pathway is thought to be a good technique for distinguishing high and low risk patients in a post-diagnosis pre-therapeutic situation to assess the success rate of a therapy. However, the pattern in this protein concentration profile is not always consistent, partially due to variance in the expression of functional paralogues and/or genetic/epigenetic changes. In this study, we found that limiting the detection of high/low risk CRC cases to a single marker protein (e.g. BclXL alone) could not be a reasonable way to solve this issue. First, we took a combined pro-and anti-apoptotic protein concentration, both of which are strongly regulated in the event of cell stress, such as tumours. We then analysed linear combinations of Bcl2 family proteins and developed a Risk Score (RS) which is a residue of the altered protein profile. RS is seen to perform the risk stratification task significantly than one of the previously suggested biomarker. We further found that the combination of patient age with expression profile enhances the performance by a significant amount.

†

† Lathwal A*, **Arora C***, Raghava GPS. Prediction of risk scores for colorectal cancer patients from the concentration of proteins involved in mitochondrial apoptotic pathway. PLoS One. 2019 ***joint first author***

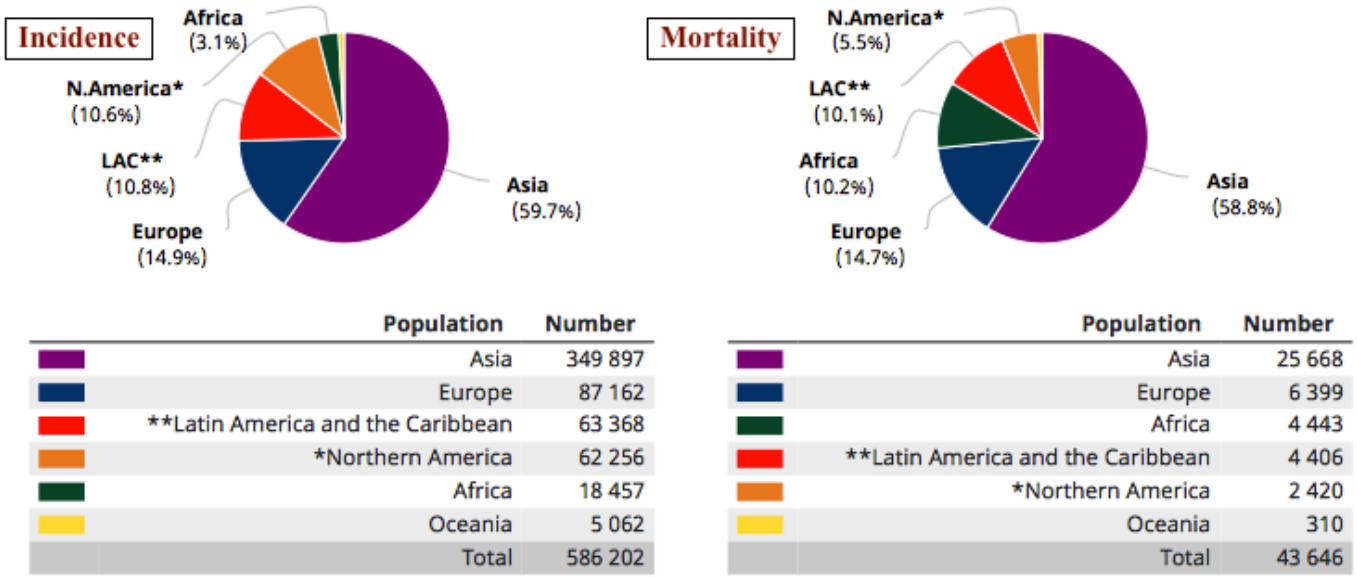
4

Risk Prediction using Gene Expression Profile

Thyroid Cancer

4.1 Introduction

The prevalence of thyroid cancer has been on a consistent rise, with the largest growth among all cancers (Mao and Xing, 2016). In 2020, thyroid cancer incidences were close to 586202 with 43646 reported deaths. **Figure 4.1** shows the distribution of incidences and deaths of thyroid cancer globally. As it can be seen, Asia has the highest number of incidences as well as deaths followed by Europe. Thyroid cancer is around three times more prominent in females and is related to increased death risk with age.



Population	Number
Asia	349 897
Europe	87 162
**Latin America and the Caribbean	63 368
*Northern America	62 256
Africa	18 457
Oceania	5 062
Total	586 202

Population	Number
Asia	25 668
Europe	6 399
Africa	4 443
**Latin America and the Caribbean	4 406
*Northern America	2 420
Oceania	310
Total	43 646

Figure 4.1 Thyroid cancer prevalence across the globe. (source: WHO-GCO)

Thyroid is a 'butterfly' shaped gland which is located near the base of the throat. There are two lobes in the thyroid gland-left and right lobes-which are separated by a thin tissue called as isthmus. The thyroid gland is responsible for secretion of hormones. Thyroid cancer can be classified into four major subtypes: i) papillary thyroid carcinoma (PTC), ii) follicular thyroid carcinoma (FTC), iii) medullary thyroid carcinoma (MTC), and iv) anaplastic thyroid carcinoma (ATC). Out of these, PTC and FTC are well differentiated tumour whereas ATC is poorly differentiated. However, PTC is the most prevalent malignant subtype, accounting for around 80-85% of all occurences of thyroid cancer (LiVolsi, 2011). PTC is generally linked with a good prognosis but indicates a bad prognosis for 20-30 percent of the patients. The existence of tumour metastases and relapses are primarily the causes of it. In certain cases it has also been shown that PTC

progresses/transforms in a more violent state, such as ATC. Due to which, there's a need for development of novel prognostic methods such that risk can be evaluated before-hand and life losses could be minimized.

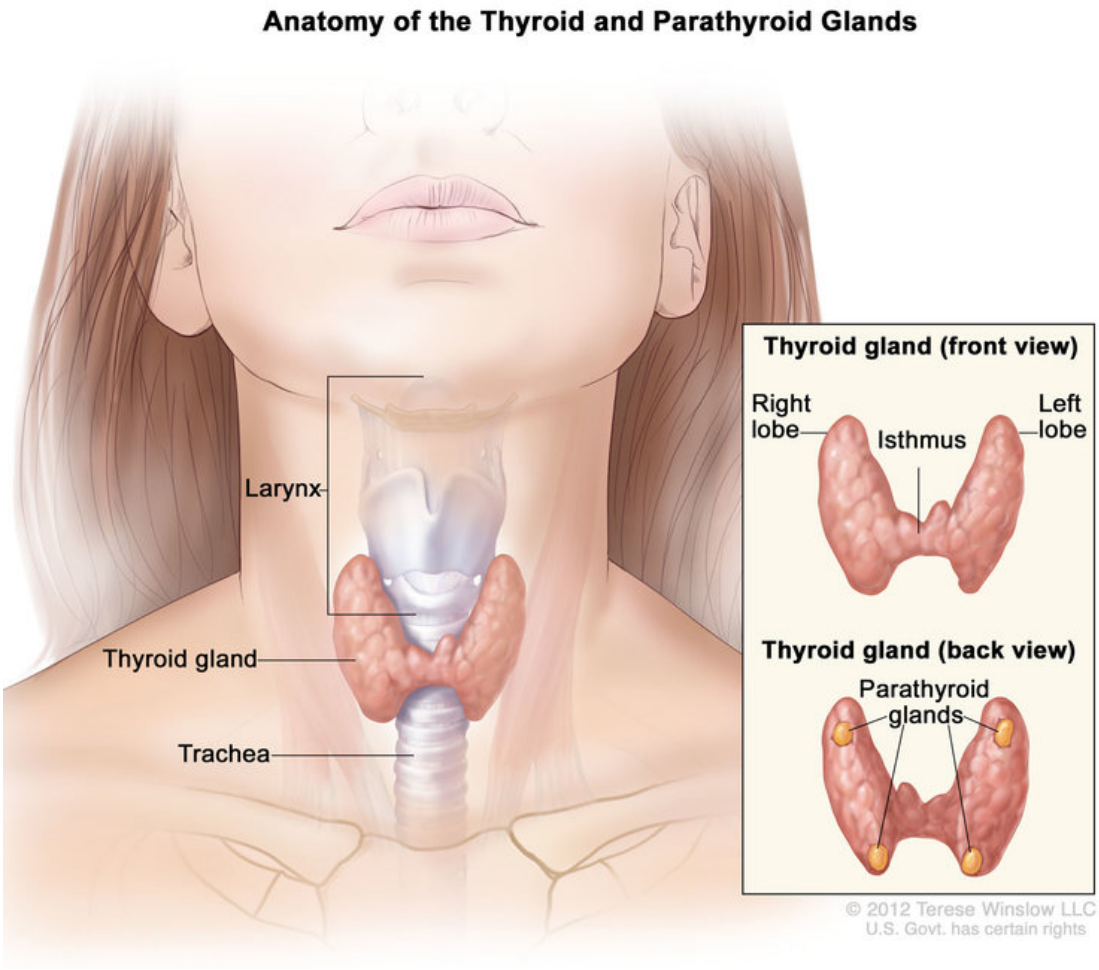


Figure 4.2 The anatomy of Thyroid gland *(Permission to use. For the National Cancer Institute © 2018 Terese Winslow LLC, U.S. Govt. has certain rights)*

It has been found that high expression of FOXE1, a member of the forkhead family, acts as a tumour suppressor in PTC. Additionally, it is reported as one of the many PTC biomarkers. In the early phase of PTC, high expression of FOXE1 was observed to negatively control PDFGA expression and hence affect PTC migration, spread and infiltration. In PTC samples, proteoglycans genes were also found to be overexpressed (Reyes *et al.*, 2019). Similarly, lower VHL gene expression has been found to be consistent with aggressive PTC and DFI characteristics (Todorovic *et al.*, 2018). Bhalla et al (Bhalla *et al.*, 2020) published about 36 transcripts of RNA

whose profiles of expression were used to identify patients with early and late-stage PTC . In addition to the above results, previous studies have recorded a number of eligible genes and biomarkers (Soares *et al.*, 2014; Bian *et al.*, 2020; Li *et al.*, 2019). The methods to accurately mine key genes from essential pathways, that can serve as prognostic biomarkers, need to be improved.

The mechanism for programmed cell-death in multicellular organisms is one such crucial process which is commonly known as "Apoptosis". Apoptosis is the process for eliminating cells in multicellular organisms. Dysregulation of apoptosis is responsible for many diseases including cancer. Numerous studies have identified key biomarkers linked with the cellular apoptosis. Charles EM *et al* present the literature related to the apoptotic molecules implicated as biomarkers in melanoma (Charles and Rehm, 2014). Another review provides extensive information related to apoptotic biomarkers such as *p53, Bcl2, Fas/FasL, TRAIL* in colorectal cancer (Zeestraten *et al.*, 2013). Several other studies have also identified key molecules with prognostic roles in other cancers like gastric cancer (Bai *et al.*, 2011; Ding *et al.*, 2020), breast cancer (Pandya *et al.*, 2020), lung cancer (Nakano *et al.*, 2020), bladder urothelial carcinoma (Zeng *et al.*, 2019), glioblastoma (Liu *et al.*, 2019) and osteosarcoma (Ma *et al.*, 2019). Apoptosis has also been found to have a crucial role in carcinogenesis of thyroid cancer. Alterations in an increasing number of apoptotic molecules such as *p53, Bcl2, Bcl-XL, Bax, p73, Fas/FasL, PPARG, TGFb and NFKb* have been associated with thyroid cancer (Wang and Baker, 2006). Since apoptotic resistance is mostly accounted for tumour proliferation and aggressiveness, apoptotic pathway has also emerged as a crucial target to develop anticancer treatments for thyroid tumours. For example, paclitaxel and manumycin are known to stimulate *p21* expression and induce apoptosis in ATC (Yang *et al.*, 2003). Lovastin inhibits protein geranylation of the *Rho* family and thus induces apoptosis in ATC (Wang *et al.*, 2001). *UCN-01* inhibits expression of *Bcl-2*, leading to apoptosis (Rinner *et al.*, 2004). Since apoptosis in PTC is a complicated multistep process involving a number of genes, it remains poorly understood and needs to be further explored at a genetic level.

Many of the current and past studies are primarily focussed on employing gene expression data for development of prognostic models. This is particularly due to the ease of extraction of expression data, as compared to protein data. In this study, we exploited the mRNA expression data obtained from The Cancer Genome Atlas-Thyroid Carcinoma (TCGA-THCA) cohort and identified key apoptotic genes that are associated with PTC prognosis. We further constructed

multiple risk stratification models using these genes and evaluated the potential of these models for prognosis using univariate and multivariate analyses, Kaplan Meier survival curves and other standard statistical tests. The 9 gene voting based model was found to perform the best and also stratified high risk clinical groups significantly. Finally, after a comprehensive prognostic comparison with other clinico-pathological factors, we developed a hybrid model which combines expression profile of nine genes with 'Age' to predict High and Low risk PTC patients with high precision. Moreover, we further validated the expression patterns of the prognostic genes by GEPIA and HPA database respectively and also verified their important biological processes. We also catalogued candidate small molecules that can modulate the expression of these genes and could be potentially employed in efficient treatment of PTC patients.

4.2 Materials and Methods

4.2.1 Dataset and pre-processing

The intial dataset comprised of RSEM normalized RNAseq values for 573 Thyroid Carcinoma samples that were retrieved in a processed data-table from 'The Cancer Genome Atlas' using TCGA Assembler-2 (Wei *et al.*, 2018) on 14th Oct 2019. The dataset, however, is open access and can also be retrieved through the TCGA-GDC portal (https://portal.gdc.cancer.gov) with the project name 'TCGA-THCA' or firebrowse (http://firebrowse.org). The list of genes involved in the apoptotic pathway were taken from previous study (Sanchez-Vega *et al.*, 2018) . Within which, data about overall survival (OS) and censoring information was accessible for 505 samples. Thus, the ultimate dataset was condensed to 505 samples, using in-house python and R-scripts, constituting RNAseq values for 165 apoptotic genes.

4.2.2 Feature selection and model development

We screened the genes related to the overall survival of the patients in TCGA datasets using univariate cox regression via 'Survival' package in R. Genes that were significantly related to the OS of the patients were selected for further analysis. Cox regression was implemented by taking the median cut-off values of the genes under consideration. BPM genes were directly correlated

with the low survival of the patients, whereas GPM genes were associated with a better outcome in the patients. Various regression models from 'sklearn package in Python were implemented to fit the gene expression values against the OS time. We also utilized different clinical features to access their contribution in predicting the OS of the PTC patients. We also implemented prognostic index-based models which were formulated as follows:

$$\mathbf{PI = \beta_1 Z_1 + \beta_2 Z_2 + ... + \beta_n Z_n}$$

Where β is the regression coefficient for any gene Z, calculated via univariate cox regression. PI was implemented to categorize patients in high and low risk groups based on best cutoff determined via cutp in 'survMisc' package. Further, voting models were also used wherein corresponding to an individual gene expression, a risk label 'High Risk' or 'Low Risk' was assigned to each patient. Thus, for *n* survival associated genes, every patient was denoted by a 'risk' vector of *n* risk labels. In gene voting based method, the patient is ultimately classified into one of the high/low risk categories based on the dominant 'label' (i.e. occurring more than at least n/2 times) in this vector.

4.2.3 Evaluation Metrics

We determined all the statistical metrics such as hazard ratio (HR), p-values, log-rank, Concordance, Wald test to evaluate the performance of the models. HR was used to assess the relative risk related to high and low risk groups. The overall workflow of the study can be found in the **Figure 4.3**.

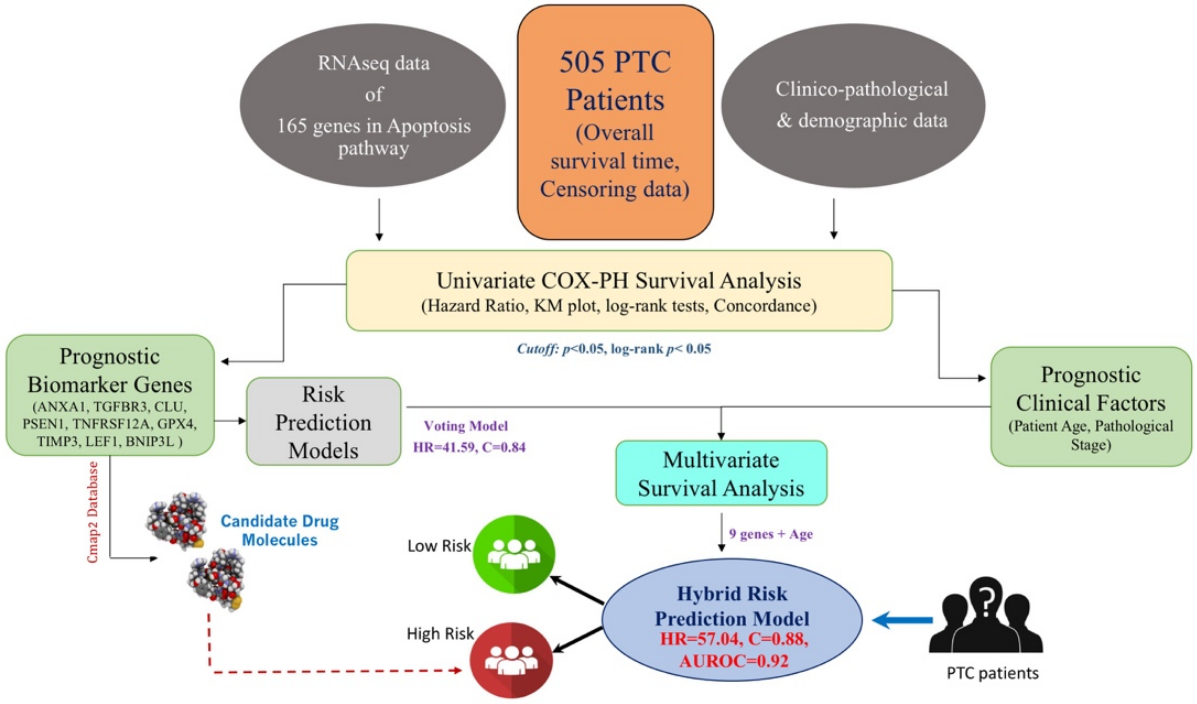


Figure 4.3 Overall workflow of the study

4.3 Results

4.3.1 Identification of prognostic biomarkers and model development

Five good prognostic marker (GPM) and four bad prognostic marker (BPM) genes were found to be associated with OS by means of univariate Cox-PH analysis. The reported GPM genes were ANXA1, CLU, PSEN1, TNFRSF12A and GPX4 while BPM genes were TGFBR3, TIMP3, LEF1 and BNIP3L. **Table 4.1** shows the results for these genes along with the metrics associated with stratification of high/low risk patients at median cutoff.

Table 4.1 The results of univariate cox regression with ">median" cutoff. Genes with HR>1" are bad prognostic markers while "HR<1" are good prognostic markers.

	Gene	HR	p-value	C	%95 CI L	%95 CI U	logrank-p
1.	ANXA1	0.14	2.82×10^{-3}	0.72	0.04	0.51	7.35×10^{-4}
2.	TGFBR3	5.68	7.90×10^{-3}	0.62	1.58	20.49	2.82×10^{-3}
3.	CLU	0.18	8.15×10^{-3}	0.53	0.05	0.64	2.92×10^{-3}
4.	PSEN1	0.15	1.20×10^{-2}	0.71	0.03	0.66	2.38×10^{-3}
5.	TNFRSF12A	0.25	1.57×10^{-2}	0.51	0.08	0.77	1.30×10^{-2}
6.	GPX4	0.27	2.98×10^{-2}	0.62	0.09	0.88	2.09×10^{-2}
7.	TIMP3	3.49	3.52×10^{-2}	0.68	1.09	11.18	2.53×10^{-2}
8.	LEF1	3.36	4.10×10^{-2}	0.68	1.05	10.77	3.00×10^{-2}
9.	BNIP3L	4.56	4.78×10^{-2}	0.68	1.01	20.46	2.05×10^{-2}

*HR: Hazard Ratio, C= Concordance Index, CI: Confidence Interval, L: Lower, U: Upper, Logrank-p: p-value for logrank test

4.3.2 Gene expression profile based risk models

Using the expression profile of nine survival related apoptotic genes, multiple risk stratification models focused on MLR, prognostic index and gene voting were built. **Table 4.2** displays the results corresponding to the performance of various models. Among these, with HR=41.59 and $p \sim 10^{-4}$ with a C-value of 0.84, the efficiency of the gene voting model was observed to be the highest. In addition, the survival curves of high/low risk classes were substantially differentiated by a voting-based model with a logrank-$p \sim 10^{-8}$. As shown in KM plot (**Figure 4.4**), the ten-year survival rate for low risk patients was approximately around 98%, which dropped to 40% for high risk patients. PI based model performed the second best with HR=17.55 and $p \sim 10^{-3}$, and regression-based RF model was the third best (and top amongst MLR models) with HR=3.09 but p-value was found to be statistically insignificant.

Table 4.2 The efficiency of various risk models constructed by leveraging nine gene expression profile.

	Model	HR	p-value	C	%95 CI L	%95 CI U	logrank-p
1.	**Voting based**	**41.59**	**3.36×10^{-4}**	**0.84**	**5.42**	**319.17**	**3.80×10^{-8}**
2.	PI	17.55	5.88×10^{-3}	0.65	2.29	134.72	6.73×10^{-5}
3.	RF	3.09	8.43×10^{-2}	0.68	0.86	11.09	5.91×10^{-2}
4.	Linear	1.59	3.98×10^{-1}	0.54	0.54	4.65	4.04×10^{-1}
5.	KNN	1.09	8.68×10^{-1}	0.56	0.38	3.12	8.68×10^{-1}
6.	Lasso	1.07	9.06×10^{-1}	0.52	0.37	3.08	9.06×10^{-1}
7.	ElasticNet	1.07	9.06×10^{-1}	0.52	0.37	3.08	9.06×10^{-1}
8.	LassoLars	1.06	9.18×10^{-1}	0.52	0.37	3.06	9.18×10^{-1}
9.	Ridge	0.84	7.43×10^{-1}	0.50	0.29	2.42	7.44×10^{-1}

*HR: Hazard Ratio, C= Concordance Index, CI: Confidence Interval, L: Lower, U: Upper, Logrank-p: p-value for logrank test, PI: Prognostic Index, RF: Random Forest, KNN: K-Nearest Neighbour

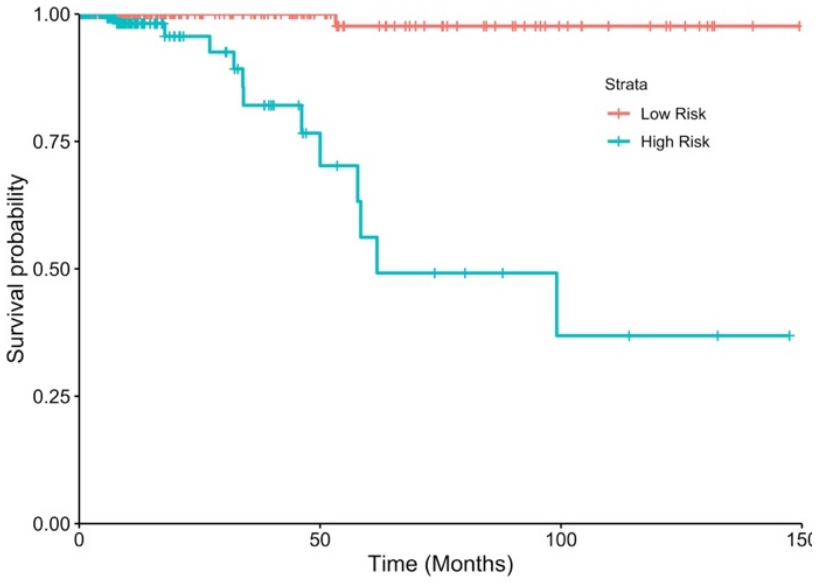


Figure 4.4 Gene voting model based risk stratification. KM plot illustrated here shows that patients with more than five "high risk" labels are at 41 fold higher risk than other patients (HR=41.59, p=3.36x10-4, C=0.84, logrank-p=3.8x10-8). High Risk: Blue, Low Risk:Red. (doi: 10.1101/2020.11.25.397547)

4.3.3 Sub-classification of patients belonging to clinical high-risk groups

In order to investigate the correlation between different clinical features and the survival of PTC patients, cox univariate regression model was implemented (**Table 4.3)**. We found that none of the clinical features were of much importance in the case of PTC patients except Age and Pathologic stage. **Figure 4.5** shows the sub-stratification by 9 gene model in the form of KM plots. A significant separation between the survival curves is seen, as denoted by logrank test's p-values.

Table 4.3 Univariate regression incorporating clinical features. "Age" is found to be the most critical factor..

Factor	Strata	N	HR	p-value	C	%95 CI		logrank-p
Age	**>60 vs <=60**	**505**	**48.65**	**1.85×10^{-4}**	**0.86**	**6.35**	**372.82**	**7.32×10^{-9}**
Pathologic Stage	**Stage III/IV vs I/II**	**503**	**9.23**	**6.61×10^{-4}**	**0.76**	**2.57**	**33.17**	**1.05×10^{-4}**
Tumour Focality	Unifocal vs Multifocal	495	5.92	8.77×10^{-2}	0.67	0.77	45.53	2.84×10^{-2}
Pathologic T stage	T3,T4 vs T1,T2	503	2.42	1.36×10^{-1}	0.66	0.76	7.75	1.17×10^{-1}
Pathologic N stage	N1 vs N0	455	1.61	4.36×10^{-1}	0.61	0.48	5.37	4.26×10^{-1}
Pathologic M stage	M1 vs M0	291	5.67	3.15×10^{-2}	0.58	1.17	27.52	7.00×10^{-2}
Race	White vs Others	413	2.20	4.49×10^{-1}	0.56	0.29	16.81	3.96×10^{-1}
Gender	Male vs Female	505	2.11	1.85×10^{-1}	0.52	0.70	6.33	2.04×10^{-1}
Laterality	Bilateral vs Unilateral	499	2.09	3.46×10^{-1}	0.49	0.45	9.63	3.85×10^{-1}
Extrathyroidal extension	Yes vs No	487	1.55	4.23×10^{-1}	0.64	0.53	4.51	4.20×10^{-1}
Residual Tumour	R1,R2 vs R0	443	3.53	4.49×10^{-2}	0.73	1.03	12.09	6.40×10^{-2}

*boldface represents statistically significant results (p-val, logrank p<0.05), HR: Hazard Ratio, C= Concordance Index, CI: Confidence Interval, L: Lower, U: Upper, logrank-p: p-value for logrank test, N: No. of Samples

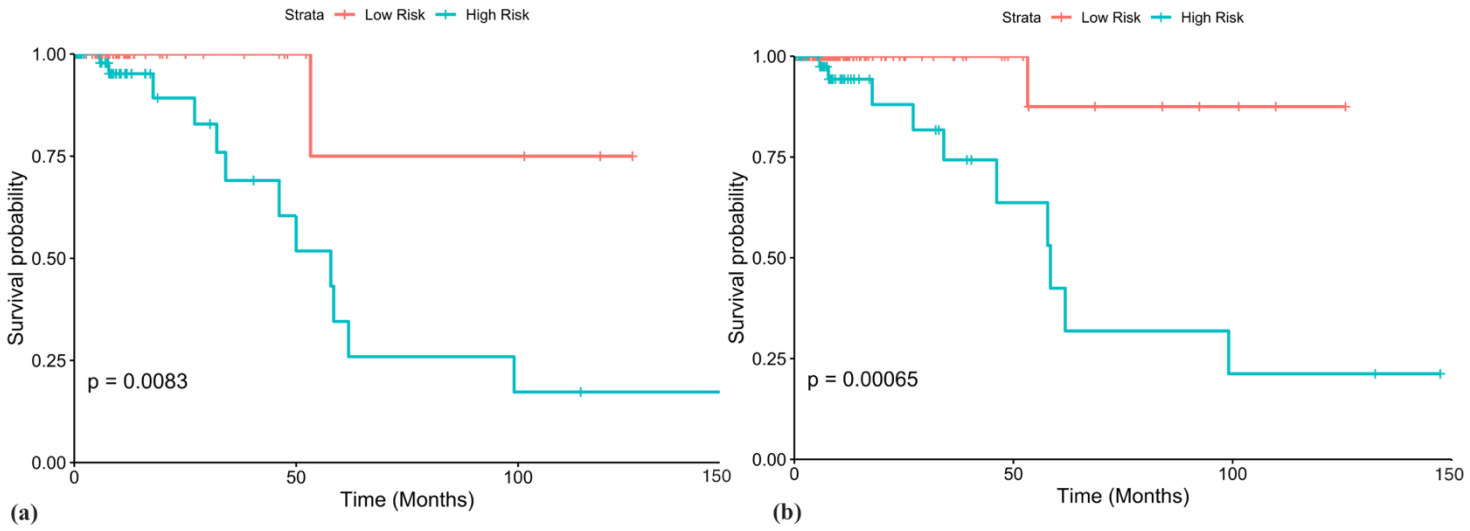


Figure 4.5 Sub-stratification of clinical "high risk" groups by voting model. (a) 113 patients whose age was greater than 60 years were segregated into "high" and "low risk" groups with an HR of 9.49, p=3.08x10^{-2} and C=0.72. (b) 167 Stage III/IV patients were segregated into "high" and "low risk" groups with an HR of 15, p=0.01 and C=0.81. p-values from logrank tests are shown in the KM plots. (doi: 10.1101/2020.11.25.397547)

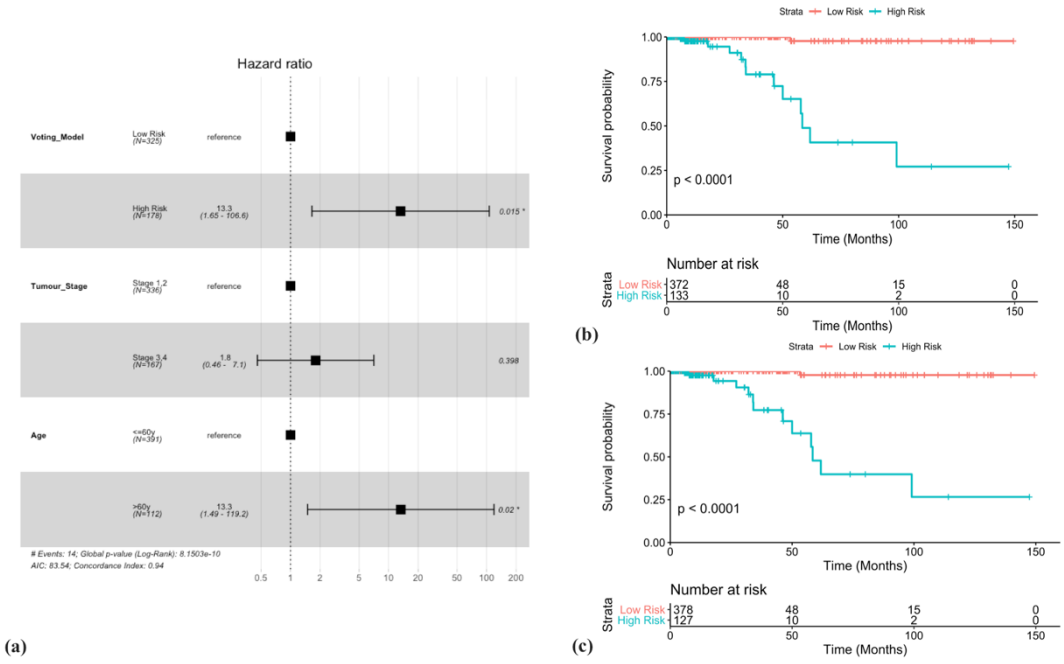


Figure 4.6 Risk stratification using hybrid models. (a) Voting model and Age were found to be independently associated covariates in a multivariate survival analysis. (b) KM plot for risk stratification by hybrid model with age cutoff of 60 years (HR=54.82, p=1.18x10^{-4}, C=0.87, %95CI: 7.14-420.90 and logrank-p=2.3x10^{-9}). (c) KM plot for risk stratification by hybrid model with age cutoff of 65 years (HR=57.04, p~10^{-4}, C=0.88, %95CI: 7.44-437.41 and logrank-p=1.4x10^{-9}) (doi: 10.1101/2020.11.25.397547)

4.3.4 Combination of age and gene voting model works best for risk-stratification

We found that patient age and gene voting model were independent covariates (**Figure 4.6a).** Next, by integrating "patient age" with the "nine-gene voting" model for risk stratification task, we established a hybrid voting model. As a result, the risk vector associated with each patient was now a 10-bit vector with 1 bit assigned to age. We found that when the age cutoff was set at 65 years (HR=57.04, C=0.88) relative to 60 years (HR=54.82, C=0.87), the model peformed better. Although there is a better distinction between the risk categories in the previous case, the 5 and 10-year survival in both models is similar.

4.4 Predictive validation

It is essential to establish that the model is not biased in the terms of data used. We therefore validated the performance of our models by a statistical approach. Using sub-samples of the entire dataset, we conducted a "predictive" evaluation of our models. For 100 iterations each, sampling sizes of 50 percent, 70 percent and 90 percent were selected. HR and C indexes corresponding to the 9-gene voting model and hybrid variants were tested for each iteration. Boxplots corresponding to the findings are seen in **Figure 4.7**. The figure reveals that the hybrid variant with an age cut-off of >65 years performs the highest in terms of HR and C values relative to other models. Consequently, an AUROC value, which denoted the classification capacity of the model, was determined. The model was seen to do well at the cut-off of 6 years out of different cut-offs used (2-10 years). A maximal AUROC value of 0.92 was achieved at this cut-off. **Figure 4.8b** reflects the ROC curve. It is important to note that this method has a limitation that same data points can be selected repetitively, thus making the results inaccurate. To avoid this, it is important to have a significant number of iterations. A benefit of using this approach, however, is that it is unbiased and can be applied to small datasets.

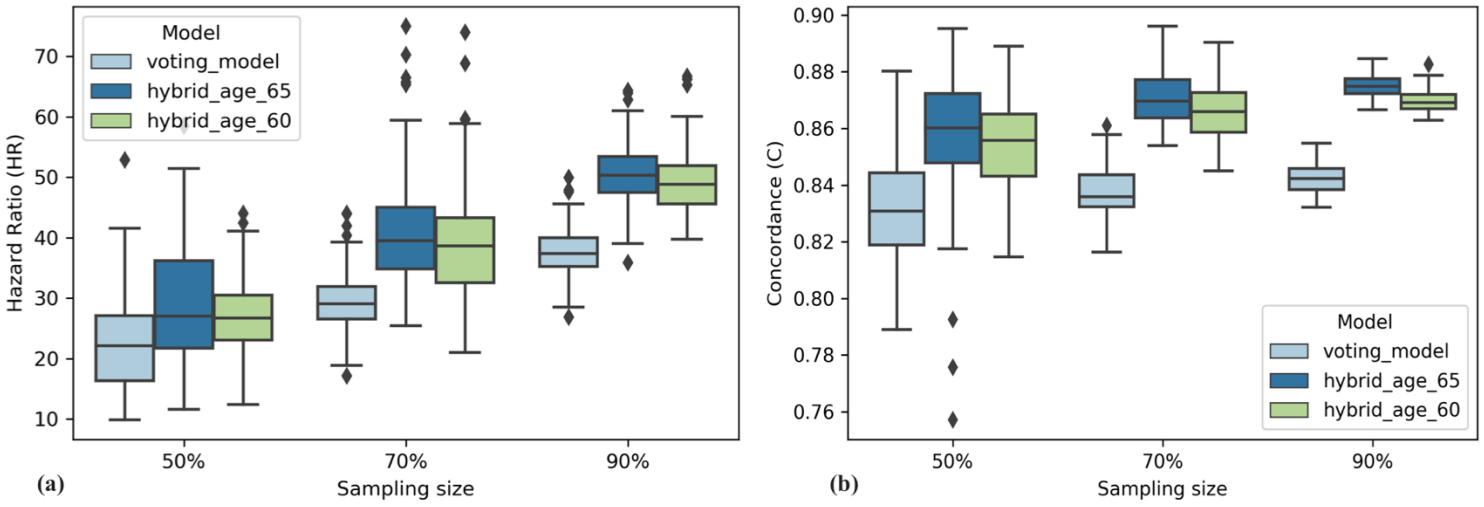
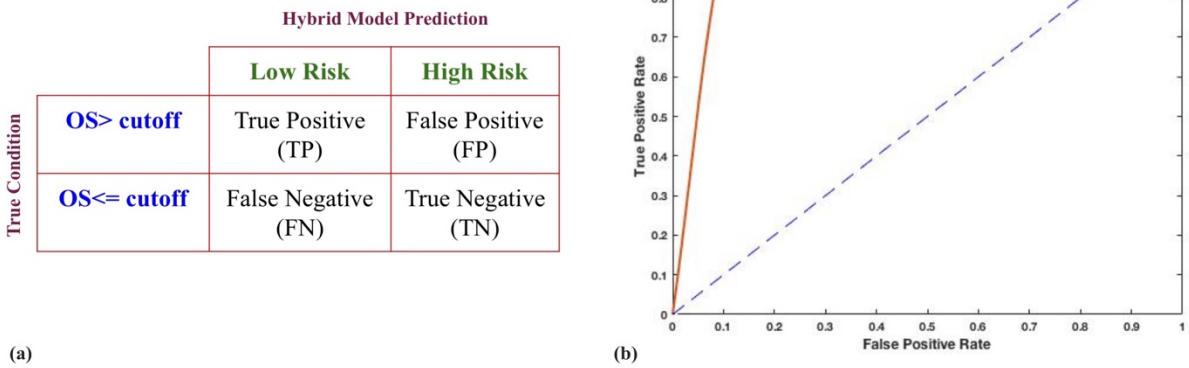


Figure 4.7 Predictive validation of voting based model and hybrid models. (a) Grouped boxplots corresponding to estimated Hazard Ratio (y-axis) for 100 iterations of data sampling (x-axis). (b) Similarly, estimation of Concordance index (y-axis) for different models using random sampling (x-axis). (doi: 10.1101/2020.11.25.397547)



Hybrid Model Prediction

True Condition	Low Risk	High Risk
OS> cutoff	True Positive (TP)	False Positive (FP)
OS<= cutoff	False Negative (FN)	True Negative (TN)

(a)

Figure 4.8 Hybrid models for classification of PTC patients using OS. (a) Terminology used for evaluation of confusion matrix. Initial risk labelling was done using an OS cutoff with patients having "OS> cutoff" labelled as positive or low risk and vice-versa for patients with "OS≤cutoff". (b) ROC curve for hybrid model using age cutoff of 65 years. AUROC of 0.92 was obtained. (doi: 10.1101/2020.11.25.397547)

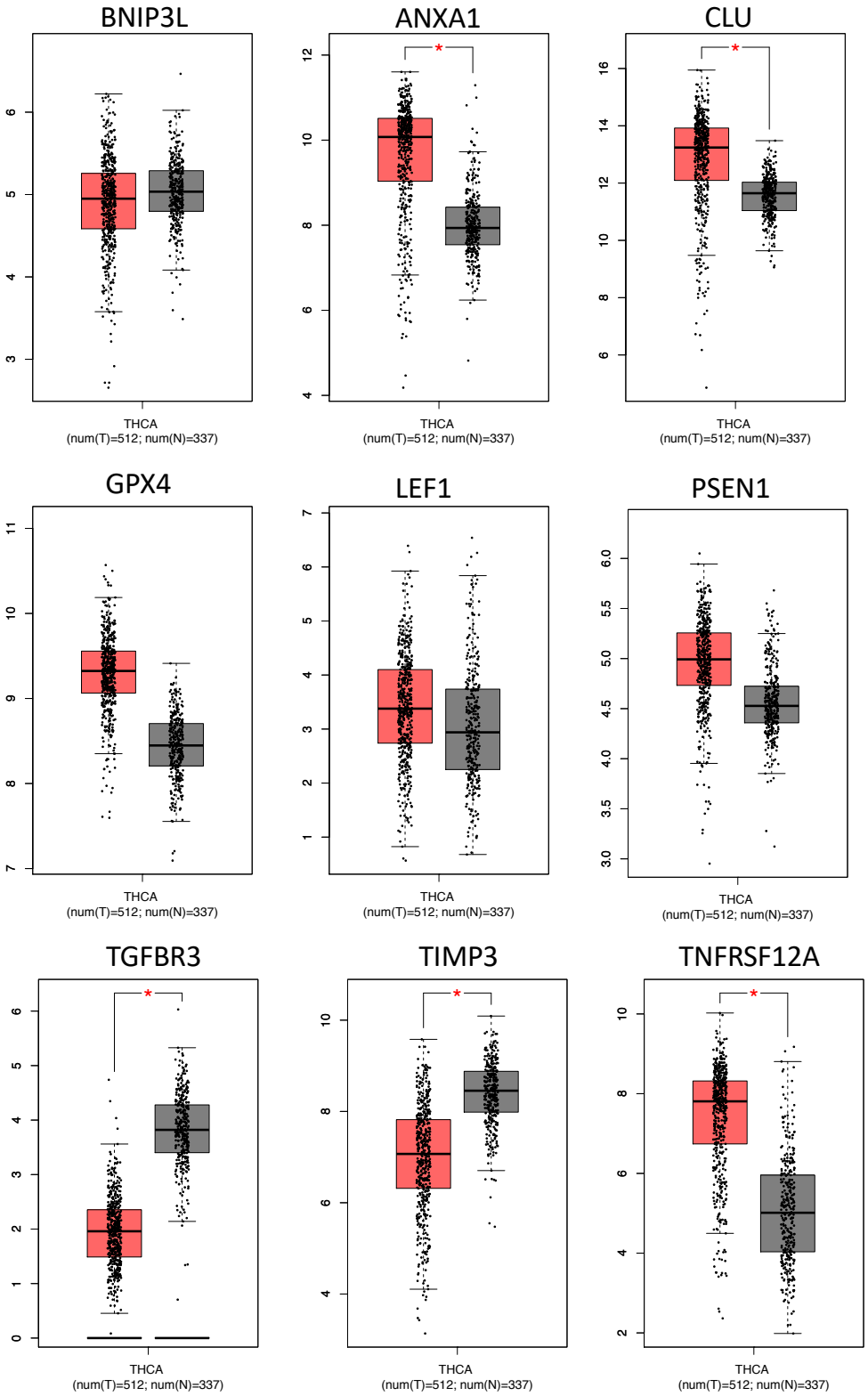


Figure 4.9 Boxplots representing the differential gene expression between normal and tumour samples on a log scale. GEPIA webserver was used to plot these by using TCGA THCA dataset. T: Tumour in red, N: Normal (TCGA,GTEX) in grey. (doi: 10.1101/2020.11.25.397547)

4.5 Validation of the Prognostic Gene Signature

With the aid of the GEPIA server, we contrasted the expression of these genes in healthy individuals (TCGA and GTEX normal samples) to patients with cancer (Tang *et al.*, 2017). Based on the results from GEPIA, it is found that the expression of ANXA1, CLU, PSEN1, TNFRSF12A and GPX4 were up-regulated in THCA, while the expression of TGFBR3 and TIMP3 were down-regulated thus elucidating their role in PTC oncogenesis (**Figure 4.9**). While, the expression of LEF1 and BNIP3L found no significant difference. Thus, it indicates that the seven genes can be considered as differentially expressed genes (DEGs) in THCA compared to normal samples.

In addition, the protein expression patterns of the prognostic genes in THCA were performed using immunostaining data available at HPA (**Figure 4.10**). The results showed that ANXA1 and PSEN1

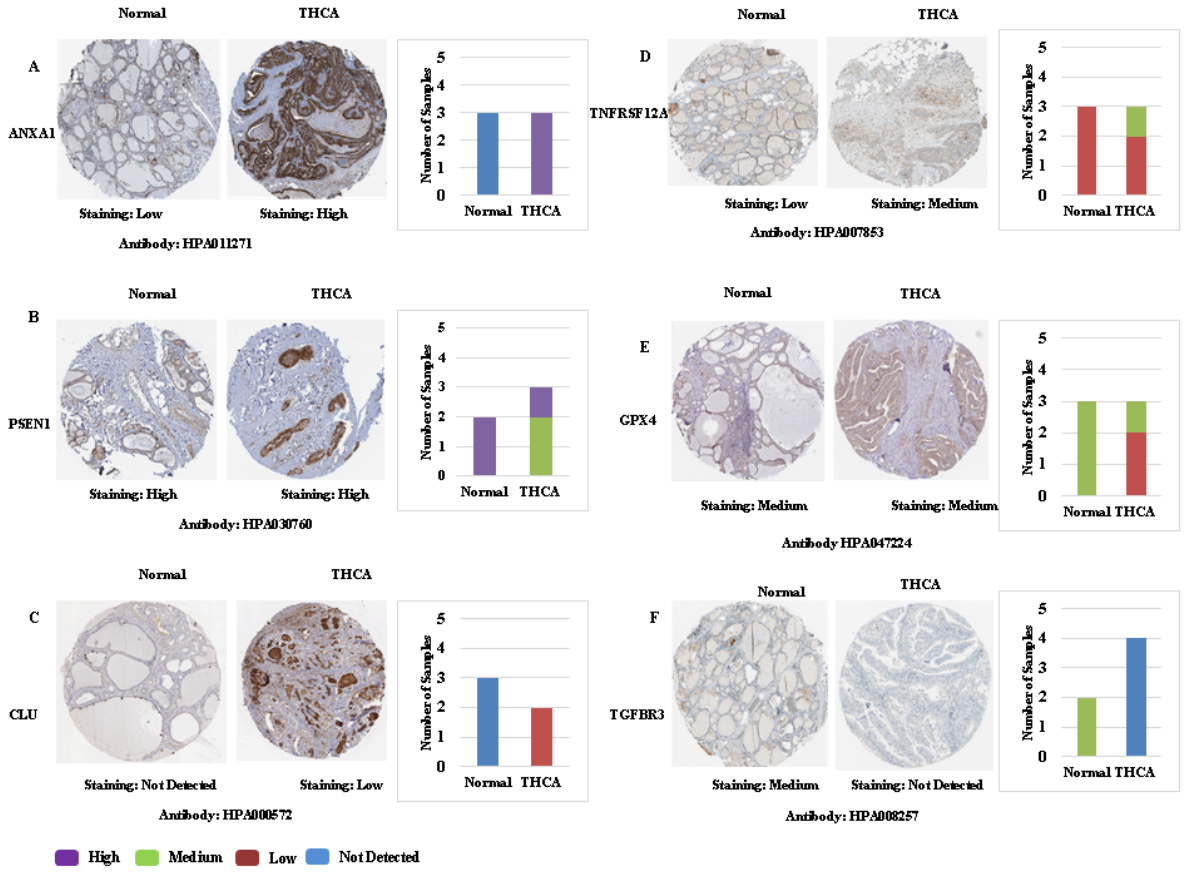


Figure 4.10 The protein expression patterns of the prognostic genes validated by HPA. (A) ANXA1, (B) PSEN1, (C) CLU, (D) TNFRSF12A, (E) GPX4, (F) TGFBR3. The staining intensity were annotated as High, Medium, Low and Not detected. The bar plots represents the number of samples with different staining intensity in HPA. (source: Human Protein Atlas, HPA)

were highly expressed in THCA. Further medium expression of GPX4 and TNFRSF12A were observed in THCA. Low expression of CLU was observed in THCA, but their expression was high at mRNA level. No expression of TGFBR3 was observed in THCA. The expression of LEF1 and BNIP3L was not detected in THCA tissues. These results validated our findings except CLU. However, the expression of TIMP3 was not recorded in HPA.

Additionally, out of these genes, *ANXA1* or annexin A1 expression has been shown to be associated with differentiation in PTC (Petrella *et al.*, 2006)**.** Western blotting experiments showed high levels of *ANXA1* in papillary thyroid carcinoma and follicular cells while undifferentiated thyroid carcinoma cells had low levels of *ANXA1* protein**.** *TGFBR3* gene was found to be differentially expressed between normal and PTC samples and was shown to be related with progression free interval (M. Wu *et al.*, 2019). The encoded *TGFBR3* protein is a membrane proteoglycan and is known to function as a co-receptor along-with other *TGF-beta* receptor superfamily members. Reduced expression of the *TGFBR3* protein has also been observed in various other cancers. *CLU* protein is a secreted chaperone which has been previously suggested to be involved in apoptosis and tumour progression. Altered *CLU* expression has also been proposed as biomarker for assessment of indeterminate thyroid nodules (Fuzio *et al.*, 2015). *PSEN1* mutations have been shown to be linked with MTC (Chang *et al.*, 2018)**.** *TNFRSF12A* was linked to aging and thyroid cancer (Lian *et al.*, 2020) and also shown to be a PTC prognostic biomarker in yet another study (Qiu *et al.*, 2018). *GPX4* is an essential seleno-protein shown to be associated with aging and cancer (McCann and Ames, 2011). *TIMP3* levels were found to be associated with *BRAF* mutations in PTC (Zarkesh *et al.*, 2018). *LEF1* expression was found to be up-regulated in PTC (Dong *et al.*, 2017) and *BNIP3L-CDH6* interaction has been linked with defunct autophagy and epithelial to mesenchymal transition (EMT) in PTC (Gugnoni *et al.*, 2017).

4.6 Therapeutic application

We found potential drug molecules using the 'Cmap2 database' (Musa *et al.*, 2018; Lamb *et al.*, 2006). As an input to 'Cmap2' a list of probe ids relating to up - regulated and down - regulated genes was used. The output consisted of a list of small molecules ranked on the basis of enrichment

scroes and p-values. Lomustine (enrichment =-0.908, p=0.0001) and Deferoxamine (enrichment = 0.663, p=0.0006) were the top 2 negative and positively enriched molecules. Lomustine is an alkylating nitrosourea compound that has been associated with the activation of apoptosis in past studies, and is already used in chemotherapy, particularly in brain tumours. (Shinwari *et al.*, 2008). Deferoxamine (DFO) is a chelator of iron that decreases the amount of iron in cells. The drug molecules could modify/change gene expression as a possible therapy in high-risk patients.

‡

‡ **Arora C**, Kaur D, Raghava GPS. Prognostic Biomarkers for Predicting Papillary Thyroid Carcinoma Patients at High Risk Using Nine Genes of Apoptotic Pathway. bioRxiv 2020.11.25.397547; doi: https://doi.org/10.1101/2020.11.25.39754 (***under review, PloS One***)

5

Risk Prediction using Clinical Features

Melanoma of the Skin

5.1 Introduction

Skin cancer is caused by the result of genomic defects in the skin cells. The resultant abnormal skin cells grow uncontrollably into a mass of tumour cells. Skin cancer often develops on sun-exposed areas of the skin such as arms, legs, face etc., however, it can also occur in other less exposed areas of the skin such as palms, beneath the fingernails or toenails. Skin cancer arises in the outermost layer of the skin i.e. epidermis. It is primarily of three types - basal cell carcinoma, squamous cell carcinoma and melanoma – related to the three types of skin cells. The topmost layer of epidermis is made up of cells called as squamous cells which form the skin lining, while the lowermost layer of epidermis is made up of basal cells which are responsible for production of new cells (**Figure 5.1**). A third type of cells known as melanocytes are found in the lower epidermis. "Melanocytes" are the class of cells which synthesize 'Melanin'- the pigment which gives the "skin" its characteristic colour. Out of different types of skin cancers, melanoma or cutaneous melanoma – which arises in melanocytes - is the most deadliest form of skin cancer (Ossio *et al.*, 2017).

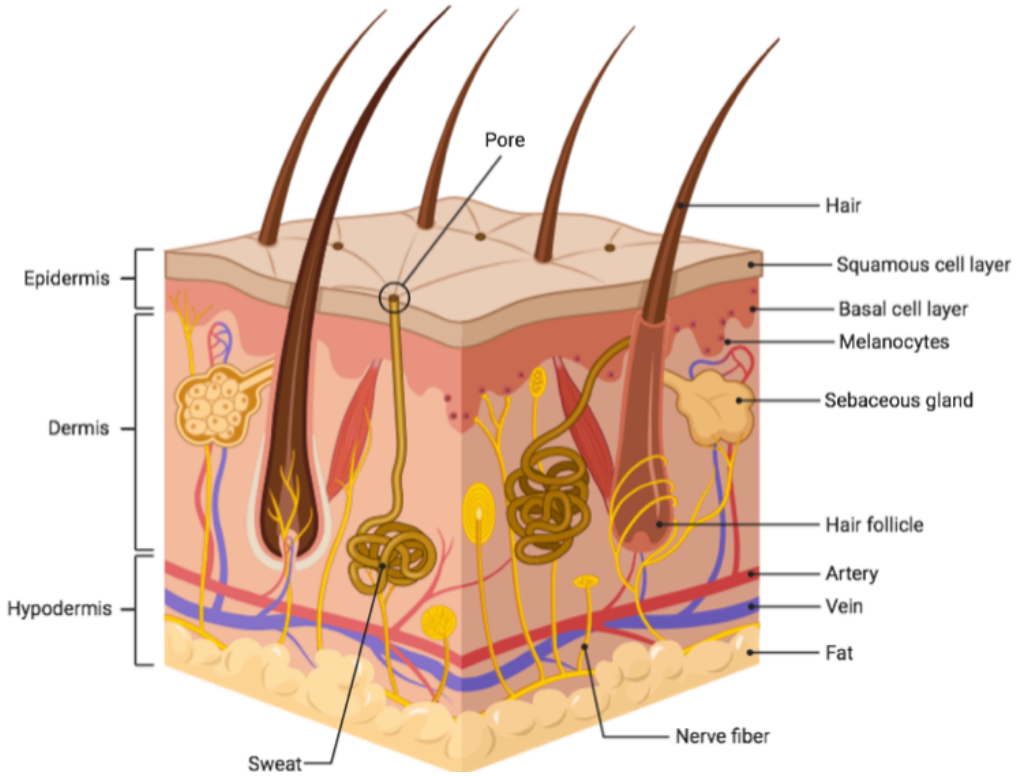
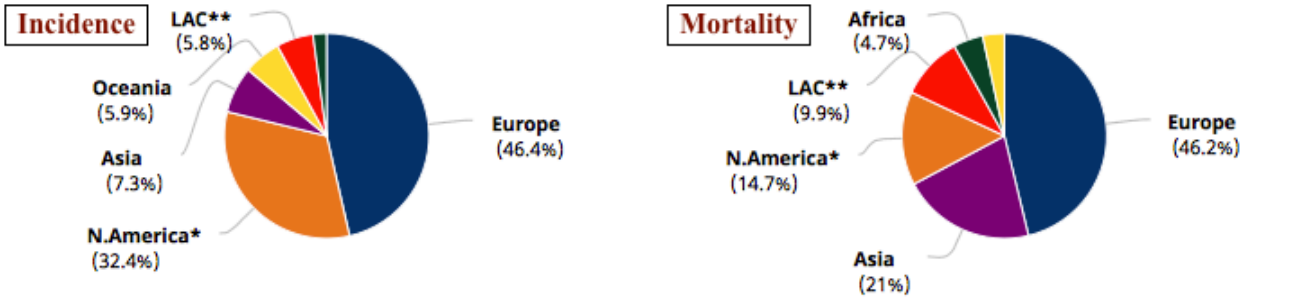


Figure 5.1 The anatomy of the skin (source: biorender.com)

The mortality rate due to melanoma has drastically increased since the last 30 years. According to the latest melanoma cancer statistics provided by the American Cancer Society (Siegel *et al.*, 2020), in US alone, around 100,350 incidences and 6,850 deaths are estimated for the year 2020. Globally, around 324,635 incidences and 57,043 deaths are estimated for the year 2020 (Global Cancer Observatory). Also, amongst the number of deaths due to melanoma, European countries account for the maximum number of deaths. The number of incidences also follow a similar pattern (**Figure 5.2**). Presently, the choice of therapy for melanoma patients is based on their segregation into different risk groups. This prognostication is performed by using AJCC TNM staging system (Balch et al. 2009) which mainly includes assessment of anatomical features from tissue samples.



Population	Number
Europe	150 627
*Northern America	105 172
Asia	23 753
Oceania	19 239
**Latin America and the Caribbean	18 881
Africa	6 963
Total	324 635

Population	Number
Europe	26 360
Asia	11 986
*Northern America	8 412
**Latin America and the Caribbean	5 657
Africa	2 679
Oceania	1 949
Total	57 043

Figure 5.2 The global incidence and mortality rates of melanoma. (Source: WHO-GCO).

The advent of high throughput sequencing techniques and availability of an explosive amount of genomic data has led to the elucidation of several underlying mechanisms associated with carcinogenesis. This insight has helped to reveal certain genes and proteins whose altered expression and/or mutation profile is utilized as potential biomarkers in some cancers. However, due to the gigantic amount of genomic data and a plethora of query molecules, identification of minimal but relevant features for risk assessment is still a challenge. It is also imperative that the novel features should complement the existing staging system and must be easily extractable for clinical feasibility. In the specific case of melanoma, a few protein candidates, including lactate

dehydrogenase (LDH), C-reactive protein and S100B, have been substantially correlated with prognostication (Gershenwald *et al.*, 2017; Deichmann *et al.*, 2004; Weide *et al.*, 2012). Of all these, only LDH for metastasis categorization has been used in the AJCC staging system so far. It is seen, however, to perform well only in patients with Stage IV disease. Another known example of multiple protein-based biomarkers-NCOA3, SPP1, and RGS1 signature-has been shown to be a major indicator of sentinel lymph node status status and disease-specific survival relative to other clinical characteristics. This 3-protein marker, while validated (Kashani-Sabet *et al.*, 2017), was also not included in the AJCC staging criteria. The most notable examples for single and multiple gene expression profile (GEP) based biomarkers include TRPM1 expression (Brozyna *et al.*, 2017), NRAS mutation status (Johnson *et al.*, 2015), BRAF mutation status (Long *et al.*, 2017), circulating miRNA biomarkers (Mumford *et al.*, 2018), DecisionDx-Melanoma (31 GEP) (Cook *et al.*, 2018), Melagenix (9 GEP) , ITLP group (Meves *et al.*, 2015) and 53-gene immune GEP (Sivendran *et al.*, 2014). The Melagenix (9 GEP) prognostic predictor was able to distinguish high and low-risk patients based on overall survival, DecisionDx-Melanoma differentiated patients based on relapse-free survival, distant metastasis-free survival and microsatellite instability. Also, the 53-gene immune GEP and ITLP are predictive models for metastasis progression and SLN positivity. However, none of the GEP based methods have been included in the AJCC staging system. We used gene expression data from over 20,000 genes in 449 melanoma patients to find GEP-based prognostic indicators in this study. We also considered genes from multiple cancer-associated pathways and created risk prediction models to examine the comparative prognostic value of apoptotic pathway genes. We compare the efficacy of GEP-based approaches to clinical factors and, as previously, attempt to construct combinatorial models. Finally, we offer a model that relies solely on clinical characteristics and outperforms GEP-based risk prediction approaches. The dataset utilized in the study comprised of gene-expression data retrieved from TCGA. The ease of extracting expression from the patients is the motivation behind using the dataset. The study's overall relevance is that it not only prioritises biological pathways important to overall survival, but it also provides a risk classification technique based on clinical features already in use. Superiority of clinical data presented here may not come as a surprise in lieu of traditional approaches, but certainly offers a topic of debate for the current emphasis on sophisticated omica based approaches. The proposed method can be used in conjunction with current staging system and be helpful in efficient management of melanoma patients.

5.2 Materials and methods

5.2.1 Dataset and pre-processing

Initial dataset including RSEM normalized RNAseq expression values for 458 patients with Skin melanoma were obtained in the form of a processed data-table from the Cancer Genome Atlas using TCGA Assembler 2 (Wei *et al.*, 2018) on 22nd May 2019. The dataset, however, is open access and can also be retrieved through the TCGA-GDC portal (https://portal.gdc.cancer.gov) with the project name 'TCGA-SKCM' or firebrowse (http://firebrowse.org. In this dataset, information on survival and censoring was available for 449 patients. Consequently, the dataset was reduced to 449 samples that had RNAseq values for 20530 genes. Following a similar approach to (Wang *et al.*, 2018), genes without expression data for more than 50% of the samples were rejected. The final dataset comprises of 449 samples with expression data corresponding to 17,292 genes. Furthermore, the final dataset was normalized using the quantile normalization method, which has been widely used in the past for similar studies (He *et al.*, 2019).

5.2.2 Identification of prognostic genes and development of risk prediction models

We collected the list of 11 cancer-related pathways and the genes associated with those pathways from a recent study (Sanchez-Vega *et al.*, 2018). Thereafter, we screened the genes related to the overall survival of the patients via Survival package in R. Risk groups were segregated on the basis of mean and median expression values of the genes, using the univariate unadjusted Cox-Proportional Hazard (Cox-PH) regression models. Genes that were significantly related to the OS of the patients were selected for further analysis. BPM and GPM genes are defined as in earlier studies. A similar screening process was implemented for all 20,530 genes. We also utilized different clinical features which includes age, gender, N staging, T staging, Breslow thickness, tumor stage etc. to access their contribution in predicting the OS of the CM patients.. Thereafter, regression models from 'caret' package were implemented to fit the gene expression values against the OS time. The fitting and test evaluations were carried using a five-fold cross-validation scheme. Hyperparameter optimization and regularization was achieved using the in-built function 'expand.grid'. The predicted OS from various regressors was used to classify high and low risk patients. We also used prognostic index (PI) based method to multiplex different gene expression

profiles together. Here, $PI=\beta_1 g_1+\beta_2 g_2+...+\beta_n g_n$; wherein β represents regression coefficient obtained for a gene g from a univariate Cox-PH model. PI was then used for risk stratification purposes.

5.2.3 Evaluation metrics

Hazard ratios were calculated to predict the risks of death associated with the high risk and low risk groups based on the overall survival time of patients. To assess survival curves of low - and high-risk groups, Kaplan-Meier (KM) plots were used. Survival tests were conducted using 'survival' and 'survminer' packages in R (V.3.4.4, The R Foundation). Utilizing log-rank tests, statistical significance was calculated between the survival curves. The assessment of the importance of the explanatory variables used in the HR measurements was done by Wald tests. The concordance index (C) showed the power of the model's predictive potential (Dyrskjot *et al.*, 2017). P-values smaller than 0.05 were deemed to be significant. The overall workflow of the study is illustrated in **Figure 5.3**.

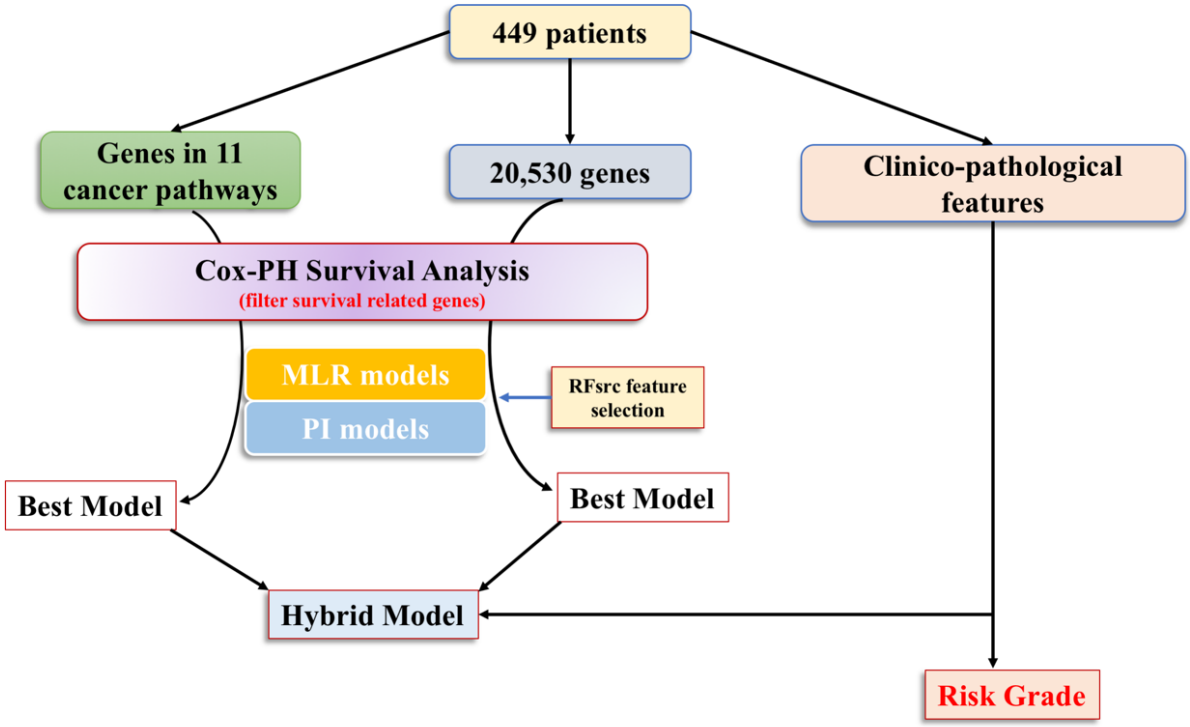


Figure 5.3 Overall workflow of the study

5.3 Results

5.3.1 Models based on genes related to cancer pathways

Amongst the 11 cancer related pathways, many have been associated with melanoma tumorigenesis. **Table 5.1** shows the PMIDs of few example studies which have explored the role of these pathways in CM progression and/or development. Combined gene count is the sum of GPM and BPM genes. The GPM, BPM and combined genesets were used for machine-learning as well as PI model development. Overall, PI models show the best results as shown in **Table 5.2.** Out of these, the combination of 29 apoptosis GPM genes with 7 NOTCH combined genes performed the best with an HR=2.57 and $p \sim 10^{-8}$. **Figure 5.4** shows the KM plots for PI for apoptotic genes and PI based on combination of Apoptosis and NOTCH genes.

Table 5.1 Genes linked to cancer-associated pathways. PMIDs are given for studies linked to the involvement of the pathways in "Melanoma" and gene count before and after univariate Cox-PH study.

S. no.	Pathway	PMID	No. of Genes	No. of GPM	No. of BPM	Combined
1	NRF2	27344172, 18353146	481	27	26	53
2	P53	32377702, 31374895	201	17	16	33
3	Apoptosis	32687246, 32645331	161	29	4	33
4	WNT	32659938, 32073511	151	7	9	16
5	CELL-CYCLE	-	128	4	17	21
6	PI3K-AKT	32626712, 32558531	105	18	11	29
7	TGF-β	31667872, 31599708	86	3	1	4
8	NOTCH	30569717, 30941830	47	3	4	7
9	MYC	32283126	25	2	2	4
10	RAS	32605090, 32568870	23	2	1	3
11	HIPPO	32407182, 32561850	22	1	2	3

*GPM: Good prognostic marker, BPM: Bad prognostic marker

Table 5.2 Risk segregation based on the prognostic index (PI). The table shows the results for each pathway and the resulting set of genes used. Patients with PI smaller than the median threshold are at lower risk than people with PI higher than the cutoff.

S. no.	Pathway	Gene set	HR	p-value	C
1	NRF2	GPM	1.87	$1.2\text{x}10^{-4}$	0.58
2	P53	Combined	2.20	$1.5\text{x}10^{-6}$	0.61
3	**Apoptosis**	**GPM**	**2.52**	$\mathbf{3.2\text{x}10^{-8}}$	**0.62**
4	WNT	GPM	1.97	$3.6\text{x}10^{-5}$	0.59
5	CELL-CYCLE	GPM	1.48	$1.6\text{x}10^{-2}$	0.57

6	PI3K-AKT	GPM	1.82	$2.4x10^{-4}$	0.58
7	TGF-β	BPM	1.48	$1.6x10^{-2}$	0.53
8	NOTCH	Combined	2.26	$9.4x10^{-7}$	0.60
9	MYC	BPM	1.67	$1.8x10^{-3}$	0.57
10	RAS	BPM	1.79	$4.5x10^{-4}$	0.56
11	HIPPO	Combined	1.67	$1.9x10^{-3}$	0.55
12	**Apoptosis+NOTCH**	**GPM+Combined**	**2.57**	**$1.5x10^{-8}$**	**0.62**

*HR: Hazard Ratio, C: Concordance Index, GPM: Good prognostic marker, BPM: Bad prognostic marker

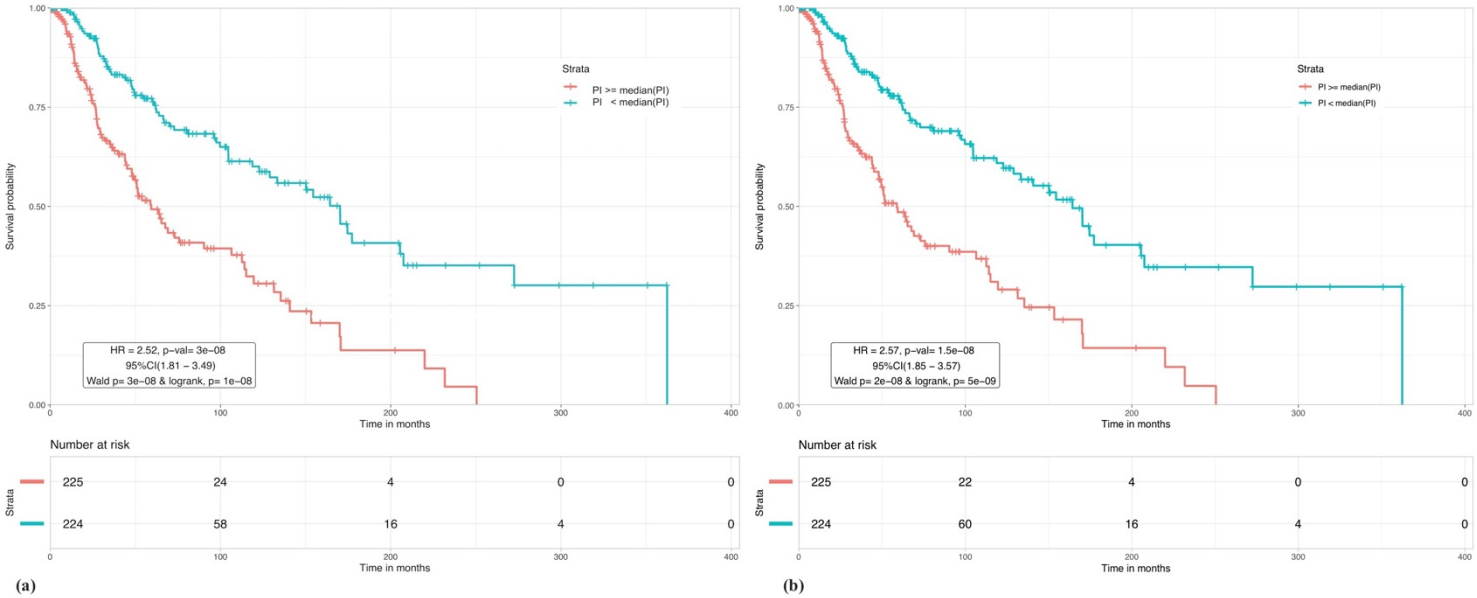


Figure 5.4 Kaplan Meier risk stratification plots of patients with CM. (a) Based on the Apoptotic Genes Prognostic Index. Patients with "PI ≥ median(PI)" are at higher risk than patients with "PI < median(PI)" with HR=2.52 and p-val=3x10-8, depending on the GPM genes. (b) Based on the prognostic index of merged genes of apoptotic GPM and NOTCH. Patients with "PI ≥ median(PI)" with HR=2.57 and p-val=1.5x10-8 are at higher risk than patients with PI < median(PI). (doi: 10.1016/j.heliyon.2020.e04811)

5.3.2 Models based on total genes

We have developed related models for the overall GPM (1343), BPM (1294) and the combined gene set (2637), in addition to developing models for pathway-specific gene sets. Feature selection was conducted on each of these three gene sets to extract the most relevant genes using random survival forests-variable hunting for 100 iterations. 58 GPM genes, 52 BPM genes and 129 combined genes resulted from rfSRC feature selection. The SVR model illustrates that HR, p-

value and concordance index have been enhanced (HR 2.77, p~10^{-9}, C 0.63). Using the chosen GPM, BPM and combined genes, prognostic index-based and MLR-based stratification was subsequently carried out. A contrast with the 52 complete BPM based models of apoptotic gene-based PI models, NOTCH gene-based regression models, apoptosis and NOTCH genes combination models.

5.3.3 Clinical-features versus GEP models

Patients have been stratified using clinical features such as AJCC pathological staging, age, TNM staging, Breslow thickness, gender and ulceration status in order to see if the models built earlier in this analysis work better than the previously identified prognostic markers. These results can be found in **Table 5.3**. While our findings align with previously recorded results, such as patients over 63 years of age, males, patients with metastasized tumours, patients with stage III/IV, etc., are at higher risk and thus display a high HR value, some of them are either marginal or have a low HR/high p-value except for Breslow thickness.

Table 5.3 Risk assessment using clinical features in CM patients. The column "N" is the number of observations for which respective information is available.

Factor	Strata	N	HR	p-value
Age	>63y vs ≤63y	449	1.83	$4x10^{-4}$
	continuous	449	1.02	$1.9x10^{-6}$
AJCC 6th ed.	Stage III,IV vs I,II	138	1.60	0.071
AJCC 7th ed.	Stage III,IV vs I,II	215	2.26	0.025
N staging	N1, N2, N3 vs N0	396	1.82	$9x10^{-4}$
T staging	T2, T3, T4 vs Tis, T1	378	1.68	$4.8x10^{-2}$
M staging	M1 vs M0	423	1.90	$9.9x10^{-2}$
Breslow thickness	>3mm vs ≤3mm	342	2.45	$3x10^{-6}$
	continuous	342	1.03	10^{-4}
Gender	Male vs Female	449	1.20	0.277
Ulceration status	Yes vs No	300	2.06	$5x10^{-4}$

*HR: Hazard Ratio, N: No. of Samples

It was observed that patients with a Breslow thickness greater than 3 mm had a 2.45 times higher mortality chance than patients with a smaller Breslow thickness. In comparison, relative to using BPM genes alone, an SVR model incorporating the 52 total BPM genes and Breslow thickness improved the output even further. At the median cutoff, which is the highest for other combinatorial models as well as previous models, an HR value of 3.19 with a p-value ~ 10^{-10} and a C value of 0.65 was achieved. This group of 52 genes was analysed for gene enrichment and the findings revealed that the following terms were enriched in various categories of GO: (i) molecular function - "catalytic activity", (ii) biological process - "cellular process" and (iii) protein class – transferase. The "pathway" enrichment of the related proteins shows that KRT4, KRT13, KRT27 and SPRR3 proteins are involved in the cornification process, which is closely associated with the risk of skin cancer (Eckhart *et al.*, 2013).

5.3.4 Superiority of Clinico-pathological features-based model

In order to incorporate the prognostic value of significant clinical characteristics, we devised a new ensemble framework. A risk point (r) was allocated to the entries corresponding to each clinical feature as r=1, 0 or -1 depending on the risk category (high risk: r=1, low risk: r=-1, unavailable: r=0), as per **Table 5.3**. Various linear combinations comprising of two or more features were evaluated and the best results were achieved with the combination of Breslow thickness, N staging, M staging and Ulceration status. We termed this combination as Risk Grade (RG) where RG for a patient is defined as:

$$RG = r(\text{Breslow thickness}) + r(\text{N staging}) + r(\text{M staging}) + r(\text{Ulceration status})$$

The hazard ratio for RG was 6.40 with a p-value of 2.49×10^{-15}. Patients with RG>1 were at higher risk than patients that had an RG ≤ 1, as represented by the KM plot in **Figure 5.5**. For high-risk cases, the 10-year mortality rate is seen to plunge to zero, while patients in the low-risk category have a 50 percent chance of survival. It should also be noticed that RG was able to stratify high (RG>1) and low risk (RG $\leq$ 1) patients with a substantial HR of 4.04 (95 percent CI 2.09-7.79) with a p-value of 3×10^{-55} even if only the patients with information available for all four clinical features were included (~259 patients) (Wald test p-val=3×10^{-5}, logrank test p-val=6×10^{-6}).

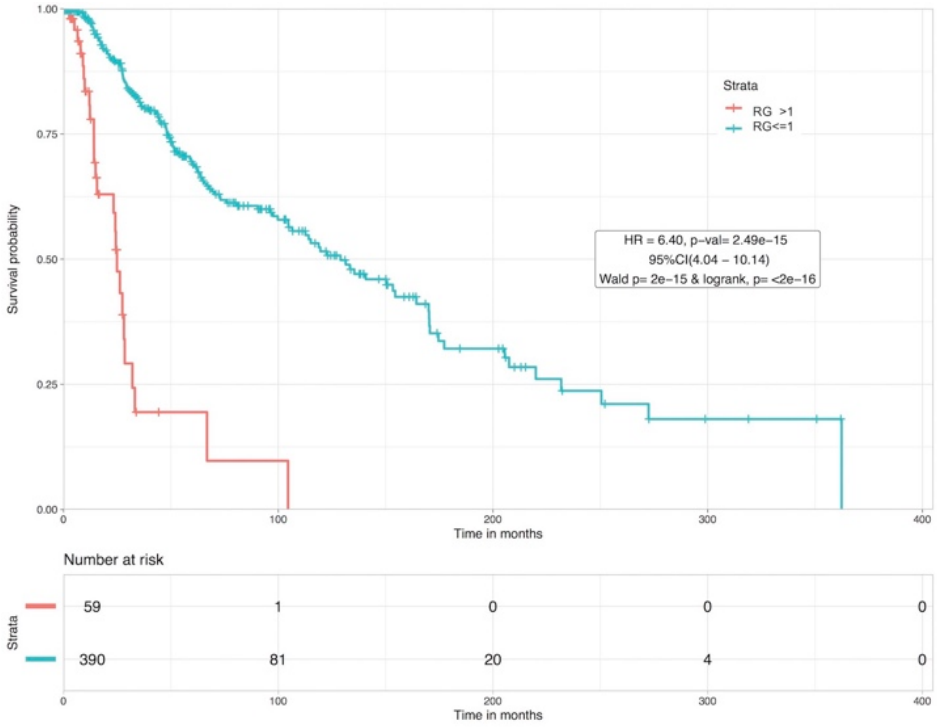


Figure 5.5 Kaplan Meier risk stratification plot based on Risk Grade for CM patients (RG). There is a greater chance of mortality for patients with "RG >1" than for patients with "RG <= 1" with HR=6.40 and p-val=2.49x10-15. (doi: 10.1016/j.heliyon.2020.e04811)

5.4 CMcrpred: web-interface and android application for risk prediction

A web server called 'CMcrpred' (http://webs.iiitd.edu.in/raghava/cmcrpred/) and Android application has been developed and is freely available in the Google Play store. Users can estimate the survival outcome and risk of a "patient with melanoma" with these facilities. A comprehensive estimate of the survival chance of a patient belonging to a given RG is given by the web server. On the other hand, for easy utility by doctors and / or patients, we have made the Android application less informative and more user-friendly. The web server was developed to configure browsing devices using a responsive HTML design.The usage for web-server and mobile application is shown in **Figure 5.6**.

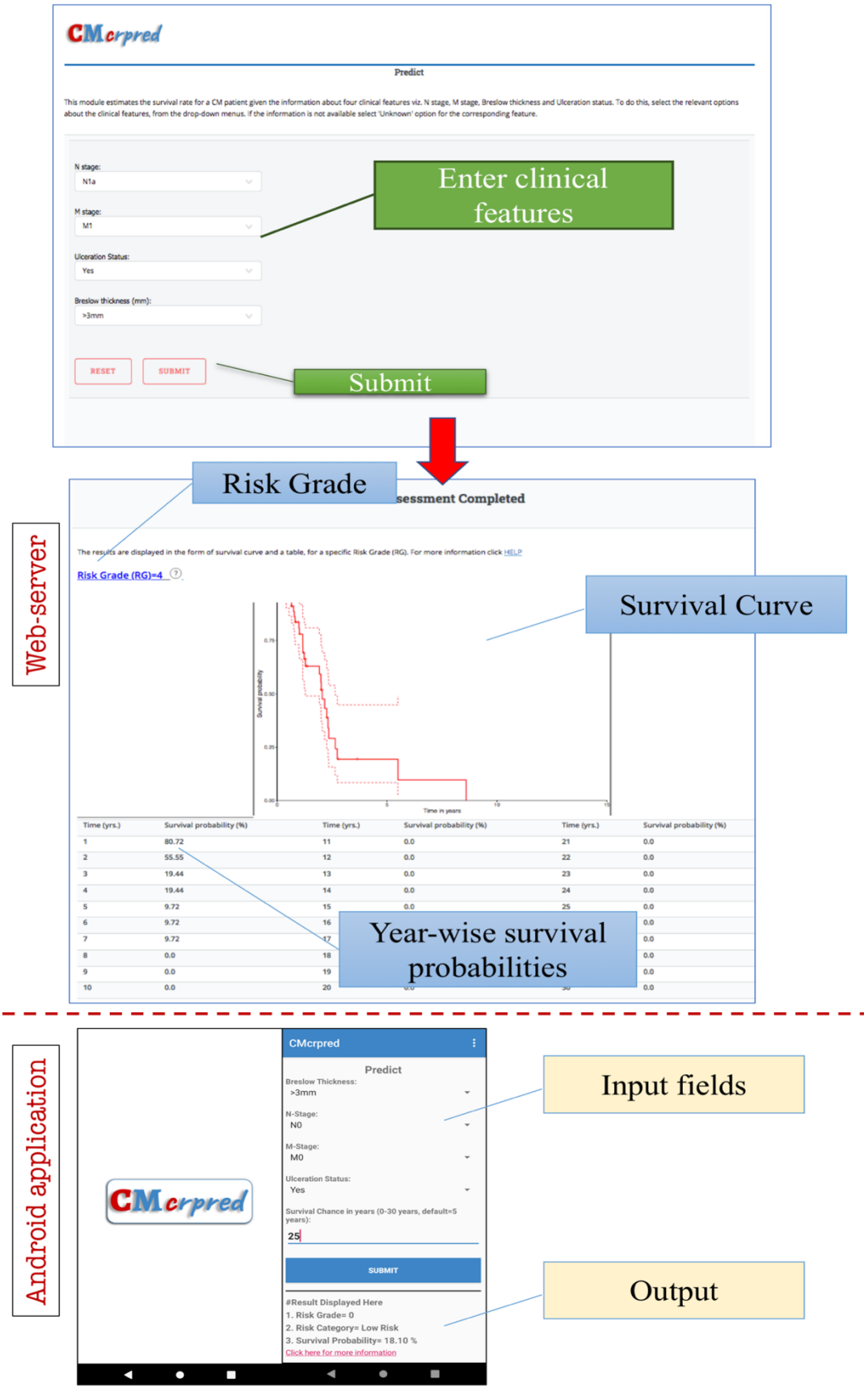


Figure 5.6 Web-server and android application functionality of CMcrpred

5.5 Comparative validation

We performed a comparative assessment of the strength of RG as prognostic marker by employing a popular melanoma survival prediction model. In order to do this, we used our dataset's features as input to the web-server 'AJCC individualised melanoma patients outcome prediction tool' (Soong *et al.*, 2010) for prediction of 5 and 10-year survival probabilities of the patient samples in our dataset. The web-server required a total of seven input features i.e whether patient had localized melanoma or regional melanoma, tumour-thickness, age, tumor burden, lesion site,

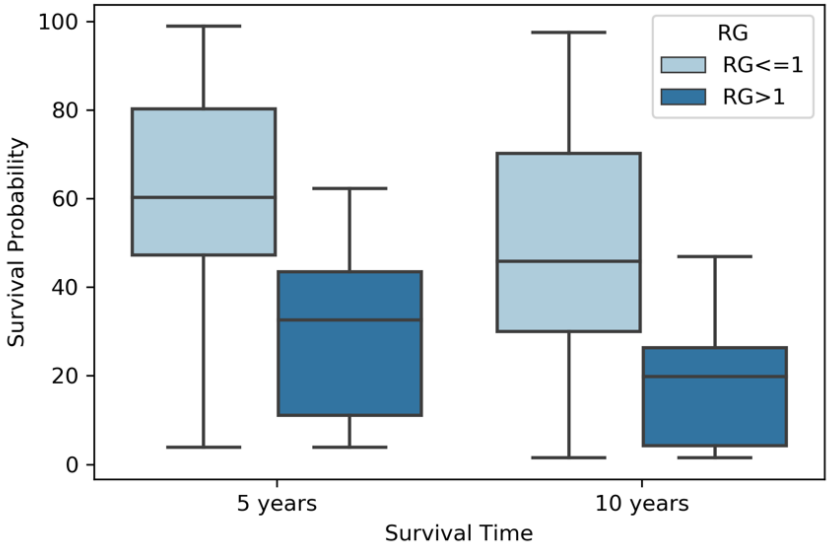


Figure 5.7 Boxplot reflecting the independent segregation of risk classes by RG based on the "AJCC individualised melanoma patients outcome prediction tool" projected 5- and 10-year survival rates. The method was used to make a total of 162 forecasts, of which 116 were low-risk patients (RG<=1) and remaining were at high-risk (RG>1). (doi: 10.1016/j.heliyon.2020.e04811)

number of nodes and ulceration status. A pre-computed RG score based on the ensemble model for was used to classify the patients based on 5 and 10-year predicted probabilities. **Figure 5.7** shows the web-server predicted survival rates between two risk groups (RG>1 or High Risk vs RG<=1 or Low Risk) in the form of a boxplot. RG was successfully able to segregate predicted low and high survival groups. Additionally, RG also requires lesser number of features than "AJCC individualised melanoma patients outcome prediction tool" for survival prediction.

5.6 Conclusion and summary

One of the main obstacles in the successful treatment of melanoma is precise risk evaluation of patients. This task is achieved by American Joint Committee on Cancer (AJCC) tumour staging system. The AJCC system is based on clinical features such as Breslow thickness, number of lymph nodes, distant metastasis etc. However, much of the emerging risk prediction approaches are based on genomic or gene-expression profile (GEP) owing to developments in technology. In this study, we sought to build novel biomarkers focused on GEP and clinico-pathological features and measured their prognostic power as opposed to current prognostic technique. Using gene expression associated with various cancer-related pathway genes, we developed risk prediction models and obtained a maximal hazard ratio (HR) of 2.52 with a p-value of ~ 10^{-8} for the apoptotic pathway. Another model improved the HR to 2.57, based on a hybrid of apoptotic and notch pathway genes. Further, we built models focusing on individual clinical characteristics and obtained a maximum HR of 2.45 for Breslow thickness with a p-value of ~ 10^{-6}. Models using the best features of clinical and gene expression data were also established and a cumulative HR of 3.19 with a p-value of ~ 10^{-9} was obtained. Finally, using clinical factors only, we established a new ensemble approach and obtained a maximum HR of 6.40 with p-value 10^{-15}. A web-based platform and an android app called 'CMcrpred' are available to promote the science community centered on this approach at (https://webs.iiitd.edu.in/raghava/cmcrpred/) and Google Play Store, respectively. This analysis shows that approaches focused on GEP-based profiles as well as commonly used AJCC staging are superseded our recent ensemble approach based on only clinical features. It also highlights the need to exploit the full potential of clinical factors for prognostication in cancer patients.

§

§ Arora C, Kaur D, Lathwal A, Raghava GPS. 2020. Risk prediction in cutaneous melanoma patients from their clinico-pathological features: superiority of clinical data over gene expression data. Heliyon 6. https://doi.org/10.1016/j.heliyon.2020.e04811

6

Gene-expression based *Universal* Prognostic Models

6.1 Introduction

Cancer is the leading cause of death worldwide and its development has been attributed to various regulatory factors (Sever and Brugge, 2015). The exploration of these regulatory mechanisms that lead to cancer has been a hot topic in recent years. Based on the exploration of these processes, there exists a plethora of biomarkers and risk prediction methods. Majority of these biomarkers/methods are specific only to a particular cancer and fail when employed for other cancers. However, with the increase in omics data, a few pan-cancer prognostic biomarkers have also been developed. Notable examples include a comprehensive analysis wherein the multi-omics data for 13 cancers was used to identify 7 genes associated with survival in 13 cancers (Zhao *et al.*, 2020), a maximum risk stratification with HR=3.03, p=0.044 in THCA patients by employing mRNA expression of Siglec-15 in 8 cancers (Li *et al.*, 2020), another study showed that the mRNA expression levels of the gene, Long intergenic non-coding RNA 1614 can be used to segregate risk groups in 11 cancers based on overall survival, with the maximum separation achieved in THCA patients with HR=4.047 and p=0.010 (Wang *et al.*, 2020). A few other studies have also elucidated the prognostic potential of genes such as WISP1 whose expression was shown to differ between cancer and adjacent normal tissues (Liao *et al.*, 2020), FUNDC1 whose expression was linked to prognosis in 8 cancers with a maximal risk separation in LIHC (Yuan *et al.*, 2019) and HSP90AA1 whose differential expression was observed in 8 cancers and was found to be a prognostic biomarker in hepatocellular carcinoma (Chen *et al.*, 2020). Apart from these, Tumor mutational burden and indel burden have also been recently shown to be linked with prognosis in 14 cancers with the best performance in CHOL (H.-X. Wu *et al.*, 2019). While these studies are promising, the challenge for finding more accurate biomarkers which offer prognostic value across a large number of cancers remains open. Since a multitude of factors cause heterogeneity of cancer, more efforts are required towards thorough investigation of cardinal molecular processes that have been associated with cancer progression and development in the past. Apoptosis is also one of the widely studied processes in the context of development of prognostic biomarkers and therapeutics which target its key components. In thyroid cancer, alterations in apoptotic molecules such as p53, BCL2, BCL-XL, BAX, p73, Fas/FasL, PPARG, TGFb and NFKb have also been associated with carcinogenesis (Wang and Baker, 2006). The downregulation of tumour suppressor gene, p53, leading to tumour development and progression is perhaps the most popular example (Bauer and

Helfand, 2006). Other examples include the downregulation of levels of pro-apoptotic BCL2 family proteins such BCL2, BCL-XL, MCL1 and upregulation of anti-apoptotic BCL2 family proteins such as BAX, BAK in cancers such as colorectal cancer, melanoma, gastric cancer etc (Frenzel *et al.*, 2009). However, the scope of these studies was limited to specific cancers with a limited set of genes/proteins. Since apoptosis consists of a large number of regulatory genes/proteins, gauging the prognostic significance of maximum number of genes/proteins involved in apoptosis across several cancers can offer a better understanding. It may also reveal several novel targets and help in development of finer biomarkers for cancer prognosis.

6.2 Materials and methods

6.2.1 Dataset and pre-processing

Normalized gene expression datasets (RSEM) and raw counts for 33 cancer cohorts were obtained from 'The Cancer Genome Atlas' (TCGA) using TCGA Assembler-2 (Wei *et al.*, 2018) in Oct 2019. The dataset, however, is open access and can also be retrieved through the TCGA-GDC portal (https://portal.gdc.cancer.gov) with the TCGA project names or firebrowse (http://firebrowse.org). A 'pan-cancer' dataset was formed by combining all the samples with raw expression values of genes across 33 cancers (Github: https://github.com/raghavagps/Chakit_Thesis). A list of 165 apoptosis genes was obtained from (Sanchez-Vega *et al.*, 2018), also available at Github. The gene expression data for these 165 genes were extracted from the downloaded TCGA cancer datasets and pan-cancer dataset. In all the datasets, only those patient samples were retained for whom overall-survival and censoring information were available. The number of samples in pan-cancer dataset was 9569 while the number of samples in each cancer cohort, N, is mentioned in **Table 6.1**. TCGA abbreviations for cancers are used.

6.2.2 Survival prediction models

Univariate unadjusted Cox proportional hazards (Cox-PH) regression models were used to screen survival-associated genes from their expression data. R packages 'survival' and 'survminer' were used to implement the Cox-PH models. Using these, Hazard ratios (HR) were computed along-with confidence intervals (%95 CI) and p-values. HR is the ratio of hazard rates representing the

death risk associated with one group as compared with another by using an appropriate cutoff of gene-expression. For comparison of survival curves between two risk groups, we used Kaplan-Meier (KM) plots and log-rank tests. Survival associated genes were identified with HR greater than or less than 1 and $p<0.05$. Concordance (C) was used to evaluate the model's predictive performance. As implemented in (Lathwal *et al.*, 2020; Arora *et al.*, 2020), Prognostic Index (PI) for n genes, g1, g2, … gn with cox coefficients β1, β2 … βn obtained from the univariate Cox-PH analyses using median cut-offs, was defined as, PI=B.g, where g=[g1 g2 g3 …. gn] and B=[β1 β2 β3 … βn]. Thereafter, risk groups were segregated by using univariate Cox-PH regression model. The cut-off value for PI was evaluated using cutp from 'survMisc' package in R. Model's performance is estimated using HR, p, %95 CI and C values. Further, for an n-gene voting model, a n-bit vector is assigned to each patient sample. Thereafter, each bit is labelled as high or low risk on the basis of corresponding classification by individual genes, using Cox-PH univariate models. Finally, the sample is allotted an overall risk label decided by majority of the labelled bits (i.e. greater than n/2 labels). The overall workflow is illustrated in **Figure 6.1**.

6.3 Results

6.3.1 Identification of prognostic biomarker genes

A univariate Cox-PH survival analysis was performed for 165 genes using each cancer's dataset. Genes were classified as good prognostic marker (GPM) or bad prognostic marker (BPM). **Table 6.1** shows the number of survival associated genes for each cancer among other details. It is seen that in most of the cancers BPM genes are more than GPM genes, showing the detrimental role of the upregulated expression of some apoptotic genes in cancer. **Table 6.1** also mentions the top genes (at most ten) for each cancer on the basis of p-values obtained from univariate survival analysis. None of the 165 genes were significantly associated with survival in 3 cancers: DLBC, TCGT and PCPG.

Table 6.1 The table shows the no. of patient samples (N), no. of BPM and GPM genes and top ten survival associated genes for 33 cancers.

Cancer	N	BPM	GPM	Total	Top Genes
LGG	511	77	17	94	WEE1,BTG3,BMP2,PLAT,SMAD7,ANXA1,PEA15,CDK2,HSPB1,SOD2
KIRC	532	50	32	82	CASP9,F2,TIMP1,IL6,CDC25B,ADD1,CCNA1,BAK1,SLC20A1,TIMP3
MESO	86	33	15	48	HMGB2,TOP2A,BRCA1,PLAT,SLC20A1,WEE1,PPP2R5B,MADD,PDCD4,LMNA
SKCM	449	10	33	43	TNFSF10,SATB1,DPYD,BIRC3,SOD2,F2R,CYLD,GCH1,CD69,PSEN2
PAAD	178	34	7	41	CASP4,TNFSF10,PSEN1,CD44,CASP2,EMP1,TOP2A,DPYD,CCND1,HMGB2
ACC	79	22	14	36	TOP2A,PEA15,BRCA1,H1F0,HMGB2,MADD,CDK2,SPTAN1,CYLD,SQSTM1
BRCA	1091	10	25	35	PTK2,NEFH,IGF2R,PLAT,DNM1L,XIAP,ETF1,NEDD9,IRF1,RARA
LAML	173	14	16	30	PDCD4,ISG20,LMNA,NEDD9,CCND2,PSEN1,HGF,SOD1,ADD1,CD44
HNSC	519	19	10	29	CCND1,BMF,CCNA1,BAK1,PSEN1,APP,TIMP1,BCAP31,SLC20A1,TNFRSF12A
UVM	80	17	12	29	ERBB3,ISG20,EREG,TIMP3,LEF1,SATB1,TXNIP,PPP2R5B,ERBB2,PTK2
CESC	304	16	10	26	EREG,CASP2,MGMT,CD2,IL1B,IGF2R,APP,NEFH,TIMP2,GCH1
KIRP	287	21	3	24	BCL2L10,TOP2A,PMAIP1,MCL1,LEF1,PPP2R5B,PEA15,DCN,IRF1,H1F0
SARC	257	7	16	23	CTH,RNASEL,GSN,IRF1,SPTAN1,CASP1,BTG2,CFLAR,TNF,CASP2
BLCA	404	7	15	22	EMP1,GCH1,HMGB2,GSTM1,CASP7,ANXA1,IFNGR1,ETF1,SLC20A1,AIFM3
LIHC	369	12	4	16	MGMT,ETF1,RARA,GPX3,EREG,CD2,DAP3,GPX4,FASLG,CDC25B
STAD	413	13	3	16	CAV1,CD44,PDGFRB,DNAJC3,EREG,TGFB2,CTNNB1,DFFA,BCL2L11,CASP6
LUSC	488	12	3	15	CD14,BTG3,EREG,CCND2,PTK2,PAK1,ADD1,HSPB1,TIMP3,SMAD7
LUAD	497	9	5	14	EREG,VDAC2,BBC3,SLC20A1,BTG2,TOP2A,RELA,CD2,GPX4,ETF1
ESCA	183	6	7	13	ENO2,IL18,TOP2A,DAP,BCL2L1,PMAIP1,ISG20,IL1A,TSPO,SATB1
COAD	297	5	5	10	BCL10,CASP4,FAS,IL6,GSR,TIMP1,BGN,LUM,ERBB2,BTG2
OV	305	4	5	9	DAP,CASP8,EMP1,BIRC3,CASP2,WEE1,PSEN1,NEDD9,SOD1
THCA	505	4	5	9	ANXA1,TGFBR3,CLU,PSEN1,TNFRSF12A,GPX4,TIMP3,LEF1,BNIP3L
KICH	65	6	2	8	IFNB1,MADD,BIK,GSR,TOP2A,PTK2,DAP3,CLU
GBM	160	6	1	7	HSPB1,FDXR,TXNIP,ANKH,EGR3,F2R,IER3
UCEC	541	5	0	5	BCL2L1,MCL1,AVPR1A,SLC20A1,ISG20
UCS	57	2	3	5	MGMT,HGF,BMF,H1F0,PTK2
CHOL	36	3	1	4	PSEN1,BNIP3L,EREG,JUN
THYM	119	2	2	4	IER3,SOD2,CD2,LEF1
PRAD	497	1	1	2	SATB1,IER3
READ	96	1	1	2	BRCA1,DNAJC3
DLBC	47	0	0	0	-
PCPG	179	0	0	0	-
TGCT	133	0	0	0	-

* N: No. of samples, BPM: Bad prognostic marker, GPM: Good prognostic marker

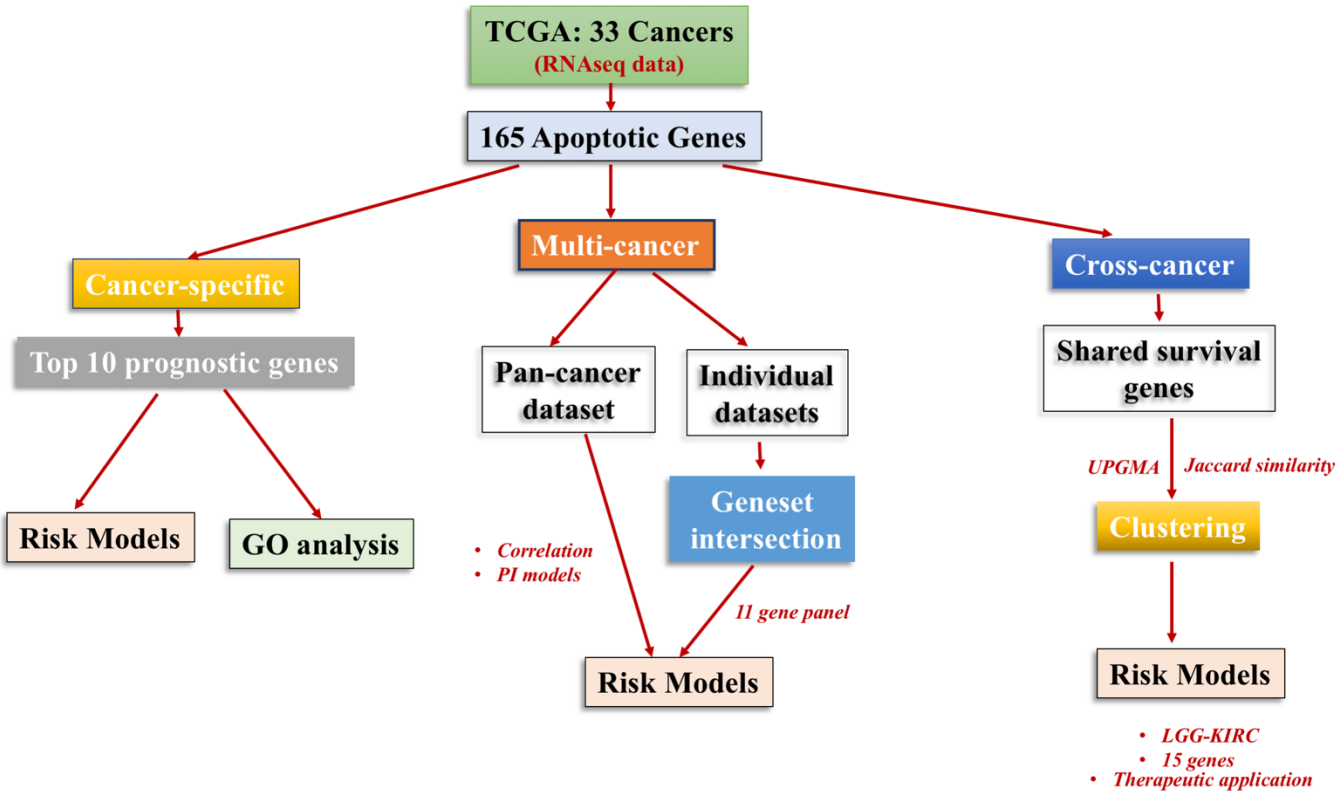


Figure 6.1 The overall workflow of the study

6.3.2 Cancer-specific prognostic models

Top genes mentioned in **Table 6.1** were used to construct models for risk stratification in 30 cancers excluding TCGT, PCPG and DLBC. Both gene voting- based models and PI models were used to segregate patients into risk groups. HR, p-values and C index were then calculated. Voting models showed the best results and are shown in **Table 6.2** (Results for PI models are not shown). For the case of PRAD and READ (2 genes each), a tie case was considered as High Risk. We also performed a GO functional enrichment for finding out the top molecular function (least p value) in the case of these cancers for top genes. **Figure 6.2a** shows the results for this. **Figure 6.2b** shows the distribution of cancers enriched to each function. We find that the molecular function 'enzyme binding' was enriched in most of the cancers viz. ACC, CESC, LUSC, SARC, STAD and UVM. Amongst these CESC and LUSC also have 'enzyme binding' as their top specific enriched function with $p \sim 10^{-5}$. There was a total of 26 genes from apoptotic pathway related to this common function. The analysis was done to see which are the underlying molecular functions

where these prognostic genes are involved in. 'enzyme binding' was the most common function amongst cancers.

Table 6.2 The performance of cancer-specific prognostic models.

Cancer	HR	p-value	logrank-p	C	%95 CI L	%95 CI U
THCA	41.59	$3.36x10^{-4}$	$3.81x10^{-8}$	0.84	5.42	319.17
UVM	40.50	$5.32x10^{-4}$	$5.12x10^{-7}$	0.85	4.99	328.82
KICH	25.61	$2.27x10^{-3}$	$3.53x10^{-5}$	0.83	3.19	205.6
ACC	22.68	$7.95x10^{-7}$	$1.63x10^{-10}$	0.81	6.57	78.31
THYM	12.53	$2.42x10^{-2}$	$6.98x10^{-3}$	0.79	1.39	112.93
UCEC	10.42	$4.51x10^{-4}$	$1.13x10^{-4}$	0.7	2.81	38.6
CHOL	8.72	$4.75x10^{-4}$	$2.45x10^{-4}$	0.77	2.59	29.4
PRAD	8.42	$4.41x10^{-3}$	$4.20x10^{-3}$	0.65	1.94	36.5
READ*	7.45	$6.50x10^{-2}$	$2.56x10^{-2}$	0.72	0.88	62.93
KIRP	5.10	$6.64x10^{-5}$	$1.27x10^{-5}$	0.72	2.29	11.37
LGG	4.99	$2.88x10^{-12}$	$1.54x10^{-13}$	0.72	3.18	7.83
CESC	4.92	$2.14x10^{-8}$	$2.98x10^{-9}$	0.71	2.82	8.6
LIHC	4.58	$7.91x10^{-11}$	$2.24x10^{-11}$	0.7	2.89	7.24
PAAD	4.41	$4.23x10^{-7}$	$1.72x10^{-7}$	0.69	2.48	7.85
COAD	4.08	$5.05x10^{-5}$	$2.42x10^{-5}$	0.67	2.07	8.05
MESO	3.99	$1.67x10^{-6}$	$2.00x10^{-6}$	0.68	2.26	7.03
KIRC	3.96	$5.41x10^{-16}$	$3.03x10^{-17}$	0.68	2.84	5.53
LAML	3.96	$3.92x10^{-12}$	$5.07x10^{-12}$	0.67	2.68	5.84
ESCA	3.80	$2.19x10^{-6}$	$3.32x10^{-6}$	0.65	2.19	6.61
UCS	3.61	$8.77x10^{-4}$	$6.13x10^{-4}$	0.68	1.69	7.67
BRCA	3.45	$2.36x10^{-9}$	$6.76x10^{-10}$	0.67	2.3	5.18
BLCA	3.41	$6.35x10^{-10}$	$3.51x10^{-10}$	0.66	2.31	5.02
STAD	3.35	$2.78x10^{-7}$	$1.39x10^{-7}$	0.64	2.11	5.31
SARC	2.81	$1.32x10^{-5}$	$1.03x10^{-5}$	0.67	1.77	4.48
LUAD	2.76	$6.94x10^{-8}$	$4.82x10^{-8}$	0.63	1.91	3.99
HNSC	2.36	$9.24x10^{-8}$	$5.80x10^{-8}$	0.62	1.72	3.24
LUSC	2.21	$1.26x10^{-6}$	$1.30x10^{-6}$	0.61	1.6	3.04
OV	2.19	$1.38x10^{-6}$	$1.16x10^{-6}$	0.61	1.59	3
GBM	2.07	$3.73x10^{-4}$	$3.22x10^{-4}$	0.61	1.38	3.09
SKCM	1.99	$2.18x10^{-5}$	$2.55x10^{-5}$	0.59	1.45	2.75

*HR: Hazard ratio, C: Concordance Index, CI: Confidence Interval, L: Lower, U: Upper

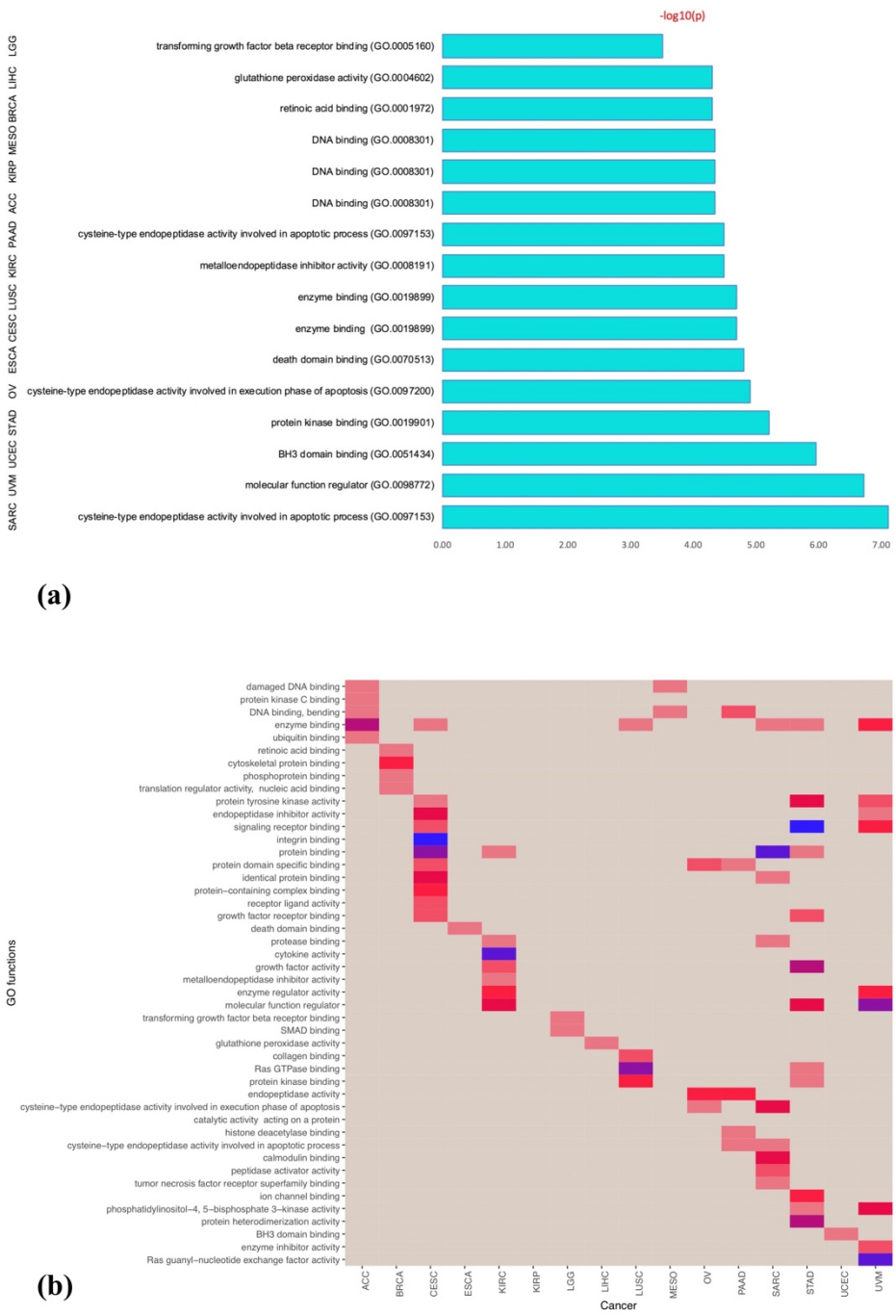


Figure 6.2 GO enrichment analysis in individual cancer cohorts. **(a)** the top enriched GO molecular function for each cancer corresponding to top genes. x-axis is the -$\log_{10}$(p-value) and y corresponds to the enriched function corresponding to the cancer. **(b)** Heatmap showing enriched GO molecular functions by top genes for each cancer. Number of genes are encoded by different colours.

6.3.3 Universal prognostic biomarkers and prognostic models

We found that there are 11 genes that play a prognostic role in more than or equal to 8 cancers (in at least 25% cancers). **Figure 6.3a** shows the role of these genes as BPM or GPM in different cancers. **Figure 6.3b** shows the 27 cancers associated with these genes. Most of the genes play a BPM role i.e. their elevated expression prevents cellular apoptosis and thus promotes tumor progression (High Risk patients). CD2 and SATB1 play a GPM role i.e. their high expression is linked with Low Risk patients. Whereas, CASP2 plays both kind of roles. Prognostic PI and voting models were constructed using the multi-cancer genes (11 gene panel) in 27 cancers. Results for voting models are shown in **Table 6.3**. This universal model performed best in UVM, THYM, PRAD, KICH and ACC based on HR and C index, where it can be readily used as a single prognostic test. Though in other cancers, the risk prediction performance of this 11 gene panel was moderate (THCA, UCEC and PAAD) to poor and thus for them, cancer specific prognostic biomarkers should be relied on for a better risk prognosis.

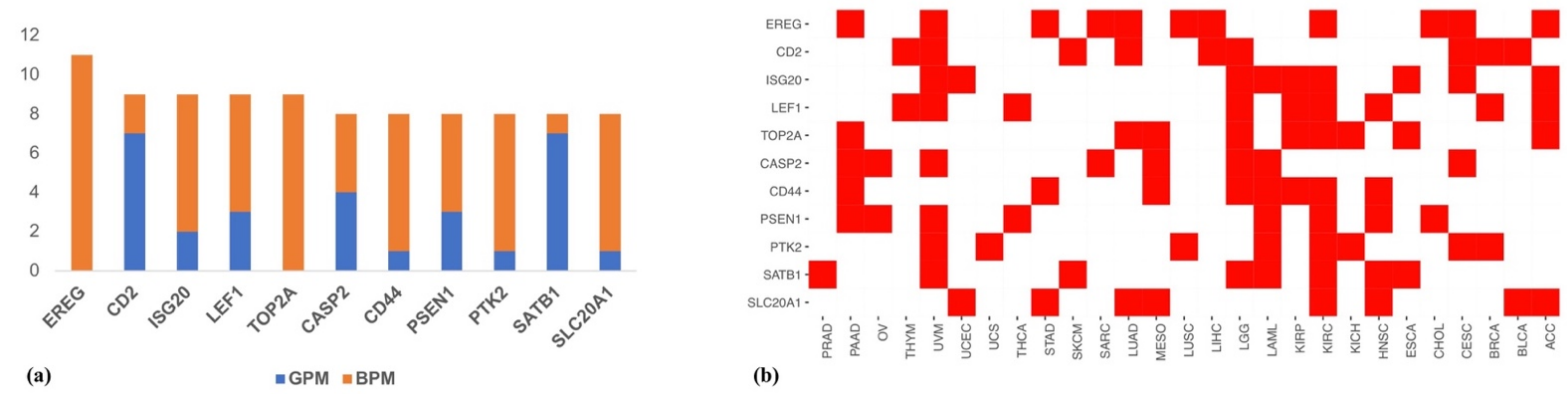


Figure 6.3 Multi-cancer survival genes. (a) Shows the distribution of role of each of these 11 genes across 27 cancers. y-axis shows the number of cancers in which the corresponding gene plays prognostic role. (b) Red blocks indicate that the gene is survival associated with the cancer.

Table 6.3 Universal prognostic model for risk prediction in 27 cancers.

Cancer	HR	p-value	logrank-p	C	%95 CI L	%95 CI U
UVM	11.74	$1.80x10^{-3}$	$1.77x10^{-4}$	0.71	2.50	55.17
THYM	10.12	$4.07x10^{-2}$	$1.48x10^{-2}$	0.77	1.10	92.91
PRAD	8.94	$4.07x10^{-2}$	$1.01x10^{-2}$	0.62	1.10	72.80
KICH	7.41	$1.27x10^{-2}$	$4.98x10^{-3}$	0.72	1.53	35.75
ACC	7.37	$3.09x10^{-5}$	$3.77x10^{-6}$	0.73	2.88	18.86
THCA	4.81	$4.94x10^{-3}$	$3.93x10^{-3}$	0.74	1.61	14.37
UCEC	4.49	$1.04x10^{-2}$	$1.07x10^{-2}$	0.64	1.42	14.18
PAAD	4.17	$4.21x10^{-6}$	$7.61x10^{-7}$	0.69	2.27	7.65
MESO	3.45	$1.29x10^{-5}$	$1.75x10^{-5}$	0.65	1.98	6.02
CHOL	3.22	$4.15x10^{-2}$	$4.89x10^{-2}$	0.65	1.05	9.91
CESC	2.93	$1.96x10^{-4}$	$8.11x10^{-5}$	0.65	1.67	5.17
KIRP	2.93	$2.85x10^{-3}$	$2.58x10^{-3}$	0.65	1.45	5.95
LIHC	2.92	$6.27x10^{-6}$	$2.38x10^{-5}$	0.61	1.83	4.65
KIRC	2.87	$1.14x10^{-9}$	$1.55x10^{-10}$	0.63	2.04	4.03
LGG	2.75	$6.37x10^{-6}$	$2.69x10^{-6}$	0.66	1.77	4.26
LUAD	2.47	$9.02x10^{-7}$	$1.09x10^{-6}$	0.63	1.72	3.54
BLCA	2.38	$2.11x10^{-5}$	$5.74x10^{-5}$	0.59	1.59	3.54
STAD	2.28	$1.06x10^{-3}$	$5.75x10^{-4}$	0.61	1.39	3.73
UCS	2.19	$4.14x10^{-2}$	$3.72x10^{-2}$	0.58	1.03	4.66
SKCM	2.07	$4.19x10^{-5}$	$9.45x10^{-5}$	0.58	1.46	2.93
ESCA	2.03	$9.00x10^{-3}$	$8.56x10^{-3}$	0.60	1.19	3.45
HNSC	1.95	$3.19x10^{-5}$	$2.49x10^{-5}$	0.60	1.42	2.67
BRCA	1.90	$2.08x10^{-3}$	$1.57x10^{-3}$	0.61	1.26	2.86
SARC	1.72	$3.16x10^{-2}$	$3.86x10^{-2}$	0.56	1.05	2.81
LAML	1.68	$6.55x10^{-3}$	$7.46x10^{-3}$	0.58	1.16	2.45
LUSC	1.59	$8.86x10^{-3}$	$1.12x10^{-2}$	0.54	1.12	2.25
OV	1.53	$1.51x10^{-2}$	$1.83x10^{-2}$	0.53	1.09	2.17

*HR: Hazard ratio, C: Concordance Index, CI: Confidence Interval, L: Lower, U: Upper

6.3.4 External validation of the universal prognostic model

The evaluation of the performance of the universal model on external cohorts is necessary for its practical translation. Therefore, we assessed the prognostic strength of the obtained eleven gene signature on various datasets. We utilized a specialized tool, SurvExpress, developed for the validation of biomarker on multiple cancer types (Aguirre-Gamboa *et al.*, 2013). SurvExpress constructed a prognostic index based model of the 11 genes that were provided. **Table 6.4** represents the result of the universal model on different cancer cohorts. The cohorts for which the expression data was unaivalable were rejected for the analysis. As observed from the results the universal model performed best for prostate cancer (HR=5.88) which is in corroboration with its performance on TCGA PAAD dataset (HR=4.49). The model is also seen to perform significantly in a variety of cancer types such as kidney cancer, ovarian cancer, colon cancer, lung cancer etc. thereby strengthening its employability as a multi-cancer risk prediction model.

Table 6.4 External validation of Universal prognostic model

S.no.	Dataset/GEO accession	HR	p-value	C	%95CI	logrank-p
1	Zhao Renal Kidney GSE3538	3.03	1.84×10^{-6}	0.69	1.92-4.79	4.82×10^{-7}
2	Tothill Bowtell Survival Ovarian GSE9891	3.97	2.17×10^{-10}	0.76	2.6-6.09	6.15×10^{-12}
3	OV-AU - ICGC Ovarian Cancer - Serous cystadenocarcinoma	2.4	6.01×10^{-4}	0.65	1.46-3.96	4.20×10^{-4}
4	Sheffer-Domany-Colon-GSE41258	3.35	5.04×10^{-8}	0.7	2.17-5.18	6.29×10^{-9}
5	Gulzar-Prostate-GSE40272	5.88	1.20×10^{-3}	0.84	2.01-17.24	1.80×10^{-4}
6	Tomida Lung GSE13213	3.97	2.41×10^{-5}	0.75	2.09-7.54	5.43×10^{-6}
7	Hoshida Golub Liver GSE10186	2.46	1.70×10^{-2}	0.65	1.17-5.15	1.40×10^{-2}
8	PACA-AU - ICGC - Pancreatic Cancer - Ductal adenocarcinoma	2.59	2.05×10^{-6}	0.67	1.75-3.84	8.67×10^{-7}
9	Peters C.Fitzgerald Esophagus GSE19417	2.8	1.00×10^{-4}	0.64	1.67-4.72	5.59×10^{-5}
10	Lenz Staudt Lymphoma GSE10846	2.68	8.09×10^{-9}	0.7	1.91-3.74	1.91×10^{-9}

*HR: Hazard ratio, C: Concordance Index, CI: Confidence Interval, logrank-p: p-value for logrank test

6.3.5 Development of cross-cancer prognostic models

It is interesting to find out which genes are shared across cancers in the context of their association with patient overall survival. The accomplishment of this task was carried out by calculating pairwise similarity between cancers c1 and c2 using Jaccard similarity index defined as:

$$J(c1,c2)=(|c1 \cap c2|)/(|c1 \cup c2|)$$

Where c1and c2 represent the set of genes that are associated with survival in cancer c1 and cancer c2, respectively. **Figure 6.4** shows the dendrograms representing hierarchical clustering plots on the basis of shared GPM genes, shared BPM genes and shared total survival genes (both BPM and GPM). Based on the Jaccard similarity index J_{all}=0.34, LGG-KIRC pair was found to be most similar in the context of survival related genes. An intersection between the set of top 20 genes (based on p-values) of both the cancers was used to develop risk stratification models.

The conjoined set consisted of 15 genes viz. BTG3, CDK2, SOD2, TOP2A, HMGB2, TIMP1, ISG20, TNFRSF12A, AFNB1, ADD1, CASP8, CDC25B, IFITM3, CD44 and GPX1. PI models were developed for both the cancers as follows:

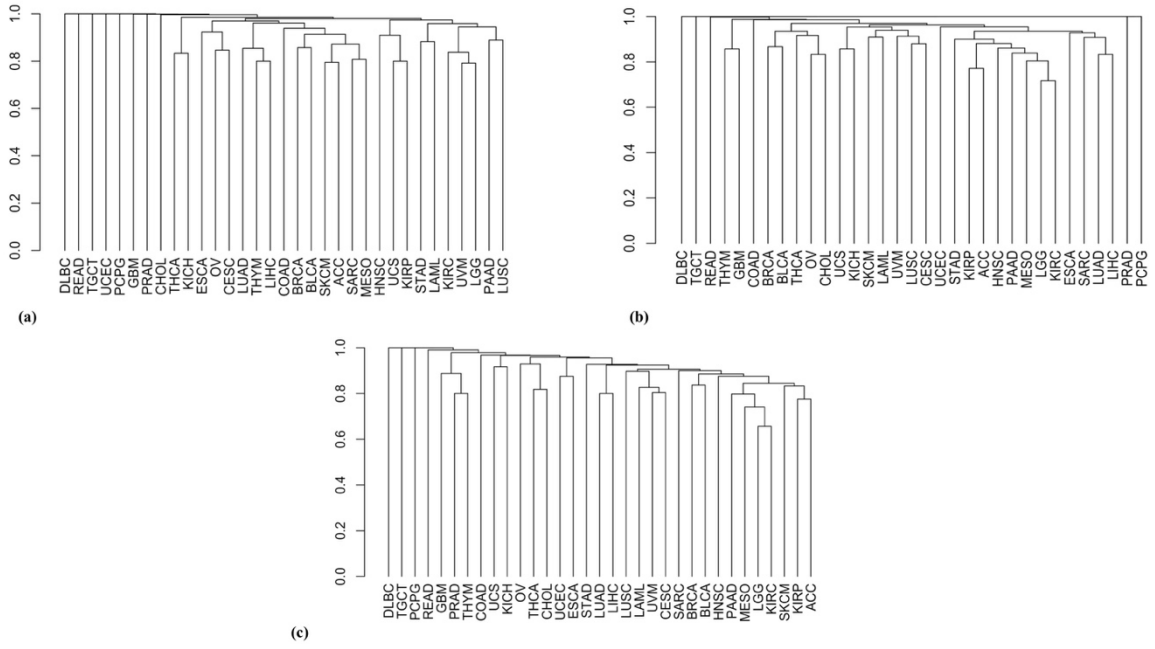


Figure 6.4 Hierarchical clustering of cancers based on (a) shared GPM genes (b) shared BPM genes and (c) all shared survival related genes.

PI_{LGG}=1.19 × BTG3 +1.07 × CDK2 +0.99 × SOD2 +1.07 × TOP2A +0.99 × HMGB2 +0.98 × TIMP1 +0.89 × ISG20 +0.91 × TNFRSF12A +0.91 × IFNB1 -0.81 × ADD1 +0.79 × CASP8 +0.77 × CDC25B +0.76 × IFITM3 +0.74 × CD44 +0.74 × GPX1

PI_{KIRC}=0.6 × BTG3 +0.7 × CDK2 +0.56 × SOD2 +0.54 × TOP2A +0.6 × HMGB2 +0.9 × TIMP1 +0.55 × ISG20 +0.61 × TNFRSF12A +0.57 × IFNB1 -0.79 × ADD1 +0.54 × CASP8 +0.82 × CDC25B +0.53 × IFITM3 +0.52 × CD44 +0.57 × GPX1

Using these, risk stratification was performed in the respective cancer as well as another cancer. While PI_{LGG} in LGG segregated the risk groups with HR=4.77, p-value=3.51x10^{-9}, C=0.68, %95CI 2.84-8.01 and logrank-p=3.41x10^{-11}; it showed a performance of HR=2.95, p-value=1.44x10^{-11}, C=0.64, %95CI 2.15-4.04 and logrank-p=1.37x10^{-11} in KIRC. Similarly, PI_{KIRC} in KIRC stratified

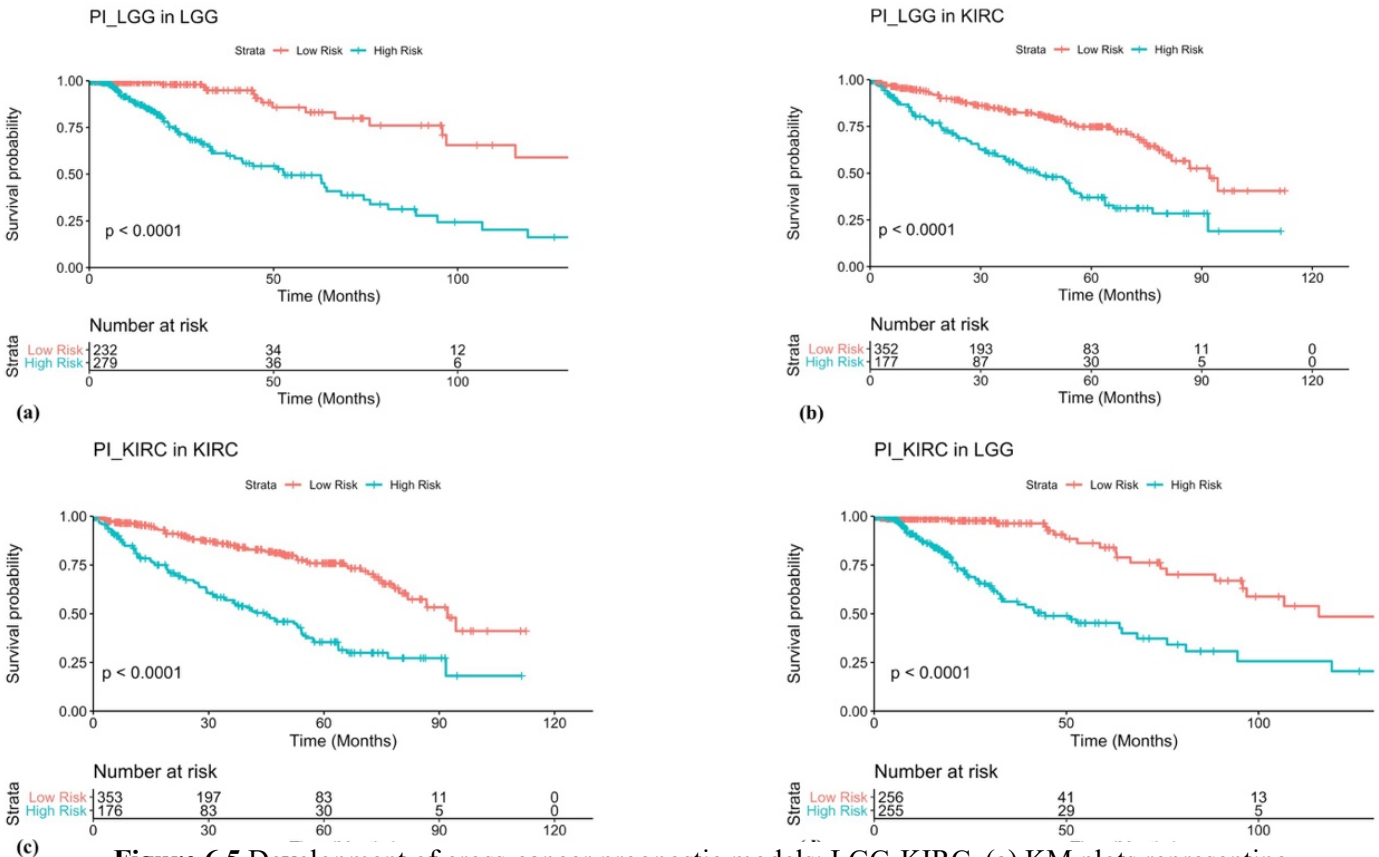


Figure 6.5 Development of cross-cancer prognostic models: LGG-KIRC. (a) KM plots representing the segregation of risk groups by PI_{LGG} in LGG cohort and in (b) KIRC cohort. (c) KM plots representing the segregation of risk groups by PI_{KIRC} in KIRC cohort and in (d) LGG cohort.

high and low risk patients with HR=3.27, p-value=1.82×10^{-13}, C=0.66, %95CI 2.39-4.49 and logrank-p=1.31×10^{-13} and in LGG with HR=4.23, p-value=1.88×10^{-9}, C=0.69, %95CI 2.64-6.77 and logrank-p=1.07×10^{-10}. KM plots corresponding to these are shown in **Figure 6.5**. It is also interesting to observe the same nature of these genes in both the cancers, as evident from the β values.

6.4 Screening of drug molecules

We further utilized the Cmap2 database and screened the potential drug molecules which could help reduce risk of death associated with high risk groups in LGG and KIRC. After querying the list of 15 genes above, we obtained the ranked therapeutic molecules. Top two enriched candidates were Genistein (enrichment=0.592, p=0) and Hexestrol (enrichment=0.918, p=0.00004). Genistein, is an isoflavone found in soy products which has recently drawn attention of the scientific community due to its potential use in treatment of cancer. Genistein is well known to induce apoptosis and prevent metastasis and has been shown to benefit colorectal and breast cancer patients (Spagnuolo *et al.*, 2015; Tuli *et al.*, 2019). Another top enriched molecule, Hexestrol, is a synthetic estrogen which was previously used for treatment of prostate and breast cancer but has been discontinued in most of the countries. However, Genistein continues to be a focus of attention in the scientific community for its anti-cancer effects.

6.5 Conclusion and summary

Numerous cancer-specific prognostic models have been developed in the past, wherein one model is applicable for only one type of cancer. In this study, an attempt has been made to identify universal or multi-cancer prognostic biomarkers and develop models for predicting survival risk across different types of cancer patients. In order to accomplish this, we gauged the prognostic role of expression of 165 apoptotic pathway genes across 33 cancers in the context of patient overall survival. Firstly, we identified specific prognostic biomarker genes for 30 cancers. The cancer-specific prognostic models achieved a minimum HR_{SKCM}=1.99 and maximum HR_{THCA}=41.59. Further, a comprehensive analysis was performed to identify universal biomarker genes across many cancers. Our best prognostic model consisted of 11 genes (TOP2A, ISG20, CD44, LEF1, CASP2, PSEN1, PTK2, SATB1, SLC20A1, EREG and CD2) and stratified risk groups across 27

cancers (maximum HR_{UVM}=11.74, minimum HR_{OV}=1.53).Further, we clustered different cancers on the basis of shared survival related apoptosis genes. This clustering approach proved helpful in development of cross-cancer prognostic models. To show the efficacy of this strategy, a prognostic model consisting of 15 genes was thereby developed for LGG-KIRC pair (HR_{KIRC}=3.27, HR_{LGG}=4.23). Additionally, we also extracted small molecules which could potentially be utilized as therapeutic candidates in LGG-KIRC high risk groups. Apart from providing a comprehensive evaluation of the prognostic potential of apoptotic genes in various cancer types, our study could be helpful in designing versatile risk management and therapeutic strategies across different cancer patients.

**

** **Arora C**, Kaur D, Raghava GPS. Universal Prognostic Biomarkers for Predicting Survival Risk of Cancer Patients from Expression Profile of Apoptotic Pathway Genes. (***under review, Wiley Proteomics***)

7

Risk Prediction using Clinical Factors

Multiple Cancers

7.1 Introduction

Conventional risk evaluation or prognostic methods in cancer care involve anatomical features derived from patient tissue samples. These features involve assessment of primary tumour characteristics and the spread to other body parts, as implemented in various staging systems such as AJCC TNM staging (Amin *et al.*, 2017). However, several other extrinsic and intrinsic factors have been widely associated with cancer risk in the past. While some of these have been included in the staging systems such as Age in thyroid cancer (Kazaure *et al.*, 2018), others are yet under scrutiny. Likewise, while ER, PR and HER2 status are now included in breast cancer staging, the intrinsic heritable risk factor associated with HBOC i.e. BRCA 1/2 mutations are only used as monitoring variables and aren't directly involved in the staging scheme. The contribution of risk

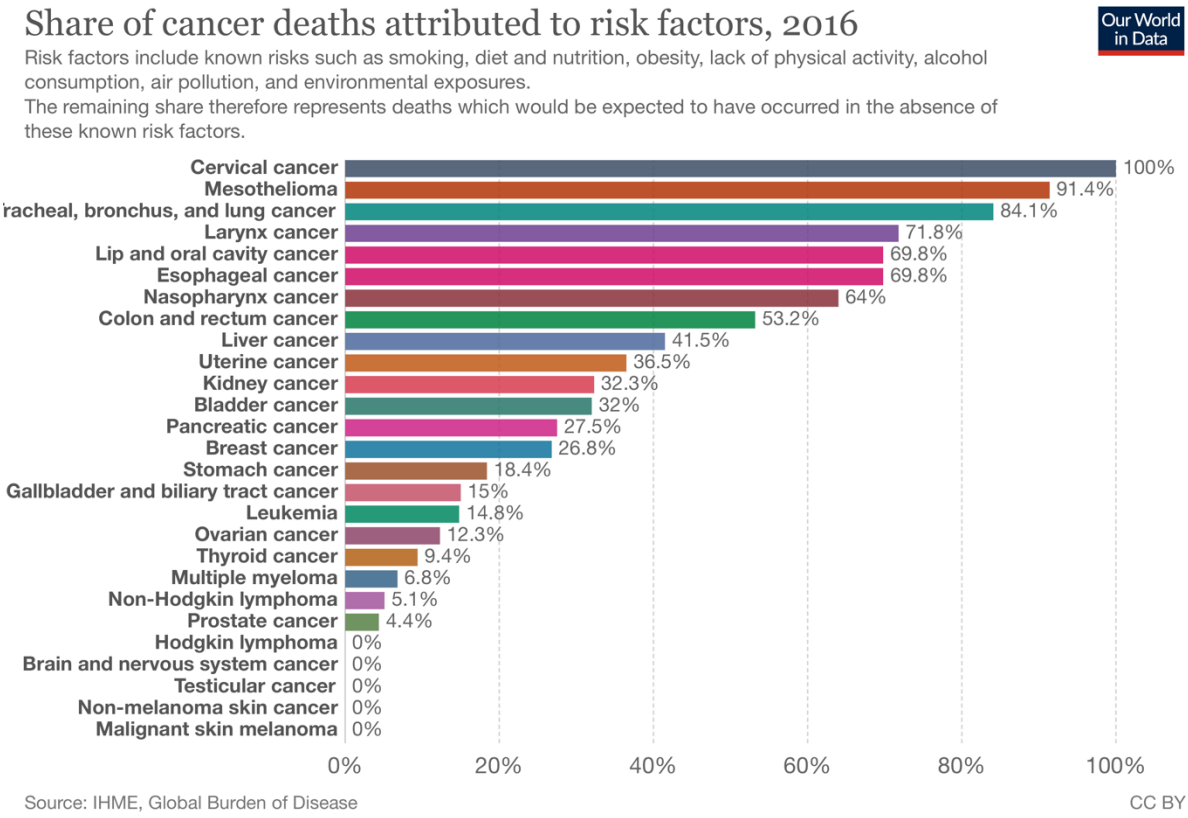


Figure 7.1 External risk factors and cancer mortality (source: ourworldindata.org)

factors versus underlying genetic factors has been a consistent matter of debate in scientific literature. According to a number of studies (Wild *et al.*, 2015; Song and Giovannucci, 2015; Ashford *et al.*, 2015), majority of the cancers are associated with extrinsic risk factors in rebuttal

to the study (Tomasetti and Vogelstein, 2015) which claimed that only one third of the cancers are caused by extrinsic factors or pre-disposed (heritable) factors. Some of the widely studied extrinsic factors associated with cancer risk are age, tobacco and alcohol consumption, lack of physical activity, dietary habits, pollution and environmental exposures. **Figure 7.1** shows the number of cancer deaths which are collectively associated to these factors with cervical cancer being the topmost. From a death toll of 5.7M in 1990 to a death toll of 8.8M in 2017, the mortality associated with cancer has shown a significant increase. The two major reasons reported for this increase is the increase in population and the increase in aged population across the world. Due to the progress in healthcare, the world has witnessed a large aging population (due to increase in average human life expectancy) and with that the number of cancer related deaths. With the increment in age, cells are supposed to lose their efficacy to fight against cancer and as a result cancer is widely termed as an old age disorder. **Figure 7.2a** shows the cancer prevalence by age in 2017. It is clear that people aged above 50 are more prone to cancer and related risk of death. **Figure 7.2b** shows this trend in death rate and the significant amount of deaths related to people with age>50. Another prominent factor attributed to risk of developing cancer is tobacco smoking. About 85% of lung cancers are caused by smoking, with an additional proportion caused by non-smokers being exposed to secondhand smoke (Warren and Cummings, 2013). The risk of lung cancer depends on the dosage, but can be significantly decreased with the cessation of tobacco use, especially if the person ceases smoking early in life. The spike in the prevalence of lung cancer follows increases in tobacco smoking in various countries across the globe. Bad clinical effects, including elevated treatment-related toxicity, increased risk of second primary cancer, reduced quality of life, and decreased mortality, are correlated with continuing tobacco use following diagnosis of cancer patients (Samet, 2013). The rate of deaths due to smoking is reported to be higher in richer countries. The GBD (Global Burden of Disease) Compare tool (https://vizhub.healthdata.org/gbd-compare/) was used to analyze the death rates due to various external risk factors such as diet and nutrition, occupational exposures, consumption of intoxicated substances, pollution, physical activity etc. The results are shown in **Figure 7.3.** As implied from the results, tobacco smoking was the biggest cause of cancer deaths followed by alcohol consumption and obesity. Although, the number of deaths attributed to these factors were less as compared to other biological alterations, they can be used in conjunction with other relevant factors such as TNM stages for a better prognostic evaluation. It is also worth mentioning that treatment procedures may also

contribute to increased death risk such as radiation exposure in radiotherapy may result into second-hand cancer or other disorders (Toma-Dasu *et al.*, 2017; Mazonakis and Damilakis, 2017),

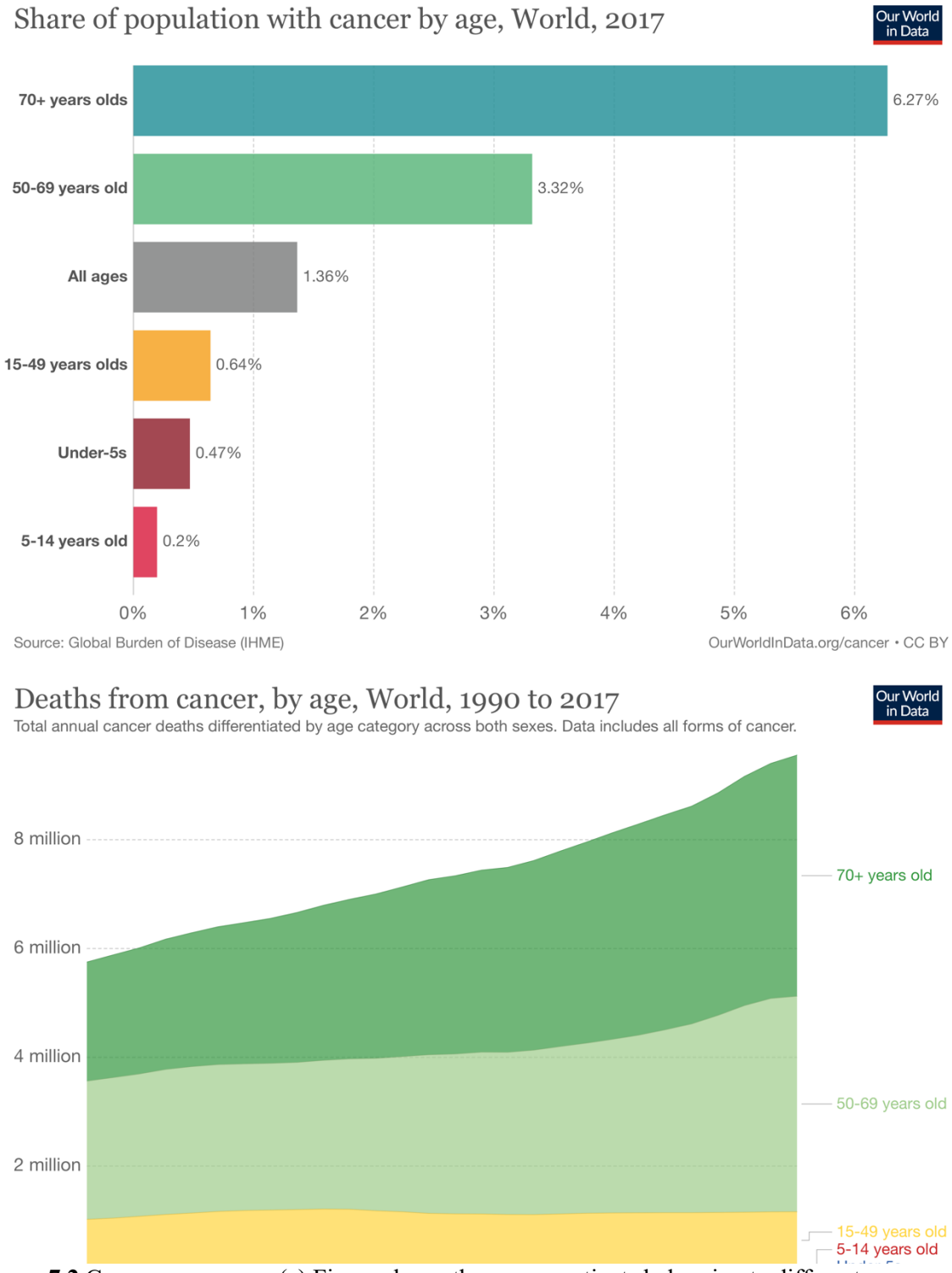


Figure 7.2 Cancer versus age. (a) Figure shows the cancer patients belonging to different age groups. (b) The increase in cancer death rates by different age groups. (source: ourworldindata.org)

exposure to toxic chemicals in chemotherapy can lead to other cancers and surgical procedures such as in old age patients or advanced cancer stage patients can aggravate their health conditions (Willaert and Ceelen, 2015).

The aim of this study is to utilize various intrinsic, extrinsic and anatomical factors to develop risk prediction models for multiple cancers. These features are collectively termed as 'clinical factors' for the purpose of our study. The patient information and registry variables were retrieved from

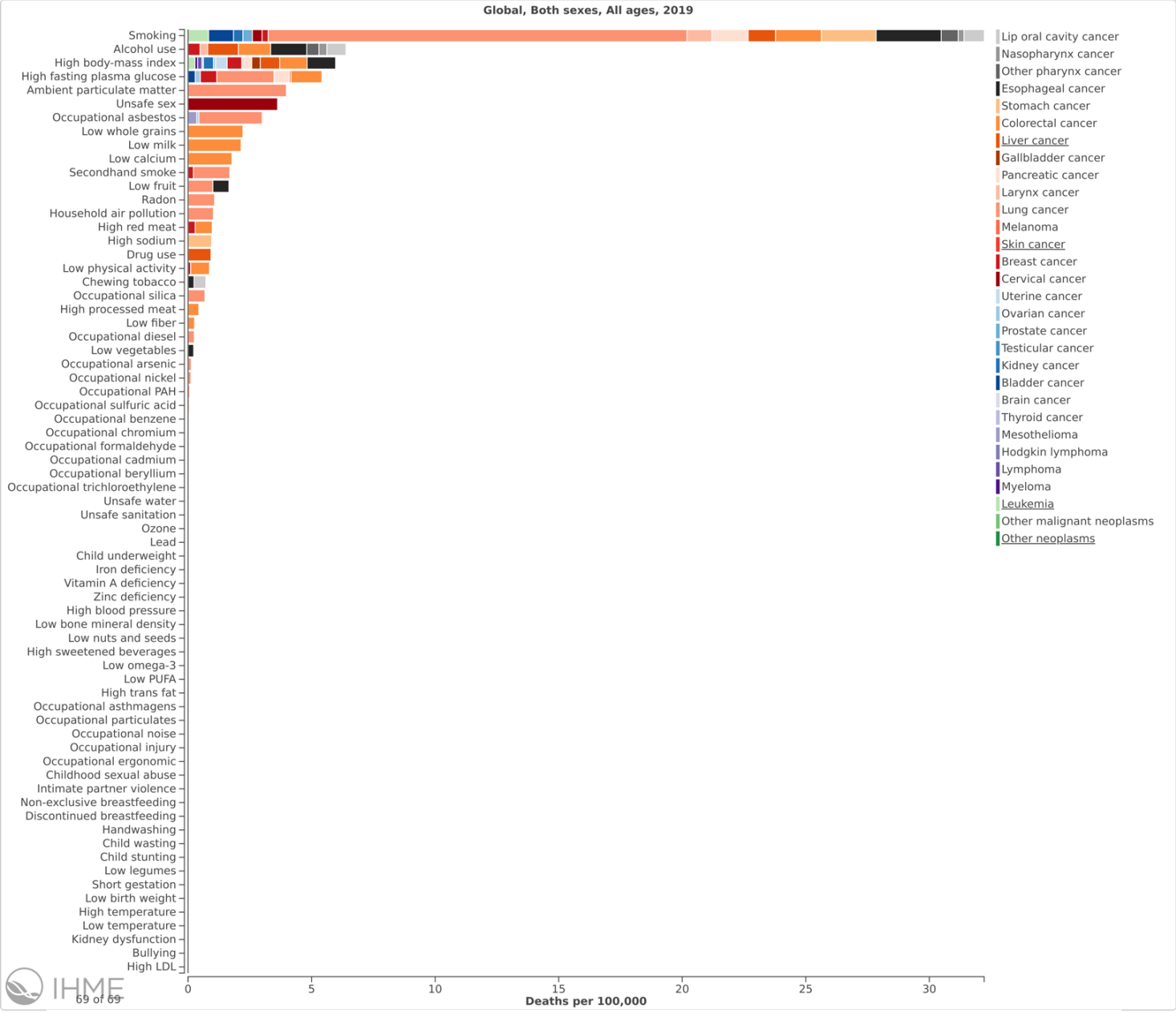


Figure 7.3 The death rates corresponding to different risk factors across multiple cancers. (source: The GBD (Global Burden of Disease) Compare tool (https://vizhub.healthdata.org/gbd-compare/)

TCGA database corresponding to 33 different cancers. Thereafter machine learning based techniques and survival analysis were used to construct models which accomplish this task.

7.2 Methods

7.2.1 Dataset

The datasets used in this study (TCGA-biospecimen) were obtained from TCGA using TCGA-Assembler 2 in Sept-2019. The datasets, however, are open access and can also be retrieved through the TCGA-GDC portal (https://portal.gdc.cancer.gov) with the TCGA project names or firebrowse (http://firebrowse.org).The datasets comprised of biospecimen data and clinico-pathological information about patients belonging to 33 types of cancer. For each cancer clinical features missing in more than half of the samples were removed. Also, samples lacking overall survival time information and censoring data were removed. Samples with survival time greater than median overall survival time were labelled as low risk and vice-versa for high risk. Each cancer-type had a large number of features, of which several were exclusive to the cancer-type while some of them were common such as age, gender etc. A table mentioning the features for each cancer-type has been given in the **Appendix A** as well as at **https://github.com/raghavagps/Chakit_Thesis**.

7.2.2 Feature selection and model development

Firstly, chi-square tests (from 'Scipy' in Python) were used to reduce the feature set corresponding to each cancer based on $p<0.05$. Decision tree classifier was used to fit the feature set and best parameters were estimated using GridSearchCV from 'sklearn' with cv=5. Thereafter, recursive feature elimination method (RFECV) from 'sklearn' was implemented to eliminate features with no additional input in machine learning model's performance. The final feature set was used for training and testing using a five-cross validation technique. Predicted labels were then used for survival analysis and stratification of risk groups. **Figure 7.4** explains the process visually.

7.2.3 Construction of risk matrices

For each cancer, a risk matrix was constructed wherein a risk probability (being at high risk) value was allocated in place of each clinical feature. This probability value for a clnical feature, *f*, was

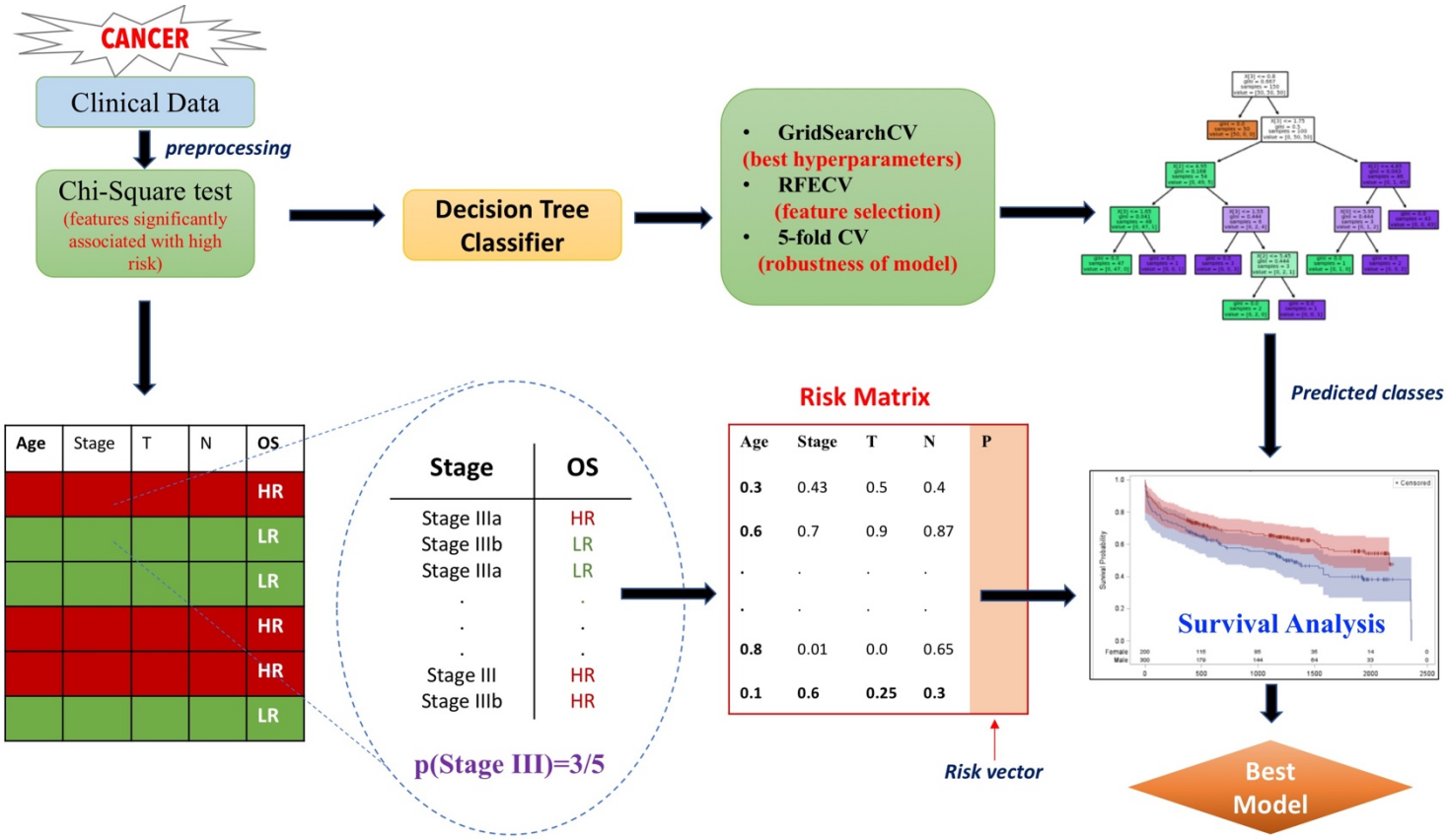


Figure 7.4 Overall design of the study

calculated using the formula : $p(f)=n_{High\ Risk}/(n_{High\ Risk} + n_{Low\ Risk})$, where $n_{High\ Risk}$ is the number of patients with feature f that are at high risk (according to OS) and $n_{Low\ Risk}$ are the number of patients with feature *f* that are at low risk. The missing/unknown features were replaced with a 50% risk probability. After the risk matrix was obtained a risk vector was created. Each element of risk vector corresponded to the mean of risk probabilities for different clinical features, for a patient. Recursive feature elimination was used for feature selection based on HR values. The construction of risk matrix and implementation for survival prediction is explained in **Figure 7.4.**

7.2.4 Survival prediction models

As implemented earlier, Univariate Cox proportional hazards (Cox-PH) regression models were used from R packages 'survival' and 'survminer'. Using these, Hazard ratios (HR) were computed

along-with confidence intervals (%95 CI), Concordance and p-values. For comparison of survival curves between two risk groups, we used Kaplan-Meier (KM) plots and log-rank tests.

7.3 Results

7.3.1 Cancer staging based prognosis

The AJCC staging information about each cancer was used to stratify the risk groups on the basis of two risk classes. High risk class contains patients with Stage 3/4 or associated substages and low risk class contains Stage 1/2 or associated substages. The pathological staging was used wherever possible, since it is considered to be more accurate. In the cases where pathological staging information was not available, clinical staging was used. **Table 7.1** shows the risk stratification using staging and the associated metrics obtained from Cox survival analysis. The highest HR was observed for THCA (HR=9.22, $p=10^{-4}$) and the lowest for CHOL, possibly due to a small data size. Cancers for which staging information wasn't available.

Table 7.1 Staging based risk stratification. cs: only clinical staging data available

Cancer type	Samples	HR	p-value	C
THCA	505	9.22	6.64E-04	0.76
ACC	90	7.82	2.32E-06	0.74
KIRP	258	4.86	7.76E-05	0.78
KICH	112	4.69	1.68E-02	0.66
UCEC (cs)	544	4.48	8.98E-07	0.70
KIRC	530	4.35	8.74E-18	0.70
UVM	79	3.33	1.18E-01	0.62
BLCA	406	3.24	4.14E-05	0.59
ESCA	166	3.09	6.24E-04	0.61
DLBC (cs)	41	3.05	3.38E-01	0.62
LUAD	495	2.78	5.75E-08	0.64
READ	161	2.72	2.37E-01	0.44
BRCA	1083	2.47	5.93E-06	0.66
STAD	411	2.39	7.05E-04	0.60
UCS (cs)	57	2.29	3.20E-02	0.62
COAD	447	2.20	6.71E-03	0.58
SKCM	414	1.60	6.29E-03	0.58

OV (cs)	580	1.60	1.27E-01	0.52
HNSC	451	1.58	3.48E-02	0.55
LIHC	351	1.48	1.17E-01	0.53
LUSC	487	1.48	3.48E-02	0.53
CESC (cs)	300	1.40	2.22E-01	0.54
TGCT	126	1.22	9.99E-01	0.75
PAAD	182	1.19	7.67E-01	0.49
HNSC (cs)	512	1.11	5.60E-01	0.51
MESO	86	1.07	8.17E-01	0.49
CHOL	45	0.71	4.87E-01	0.54

*HR: Hazard ratio, C: Concordance Index

7.3.2 Age and gender versus survival risk

We performed an exploratory analysis, to explore the role of a few common features in cancer risk. **Figure 7.5** shows the distribution of high risk patients i.e. the patients which survived less than the median overall survival time of the dataset, with respect to Age, Gender, Ethnicity and Stage. Staging, is already a validated method of risk determination, and has also been analysed in the context of survival in the previous section is seem to play an expected role with majority of

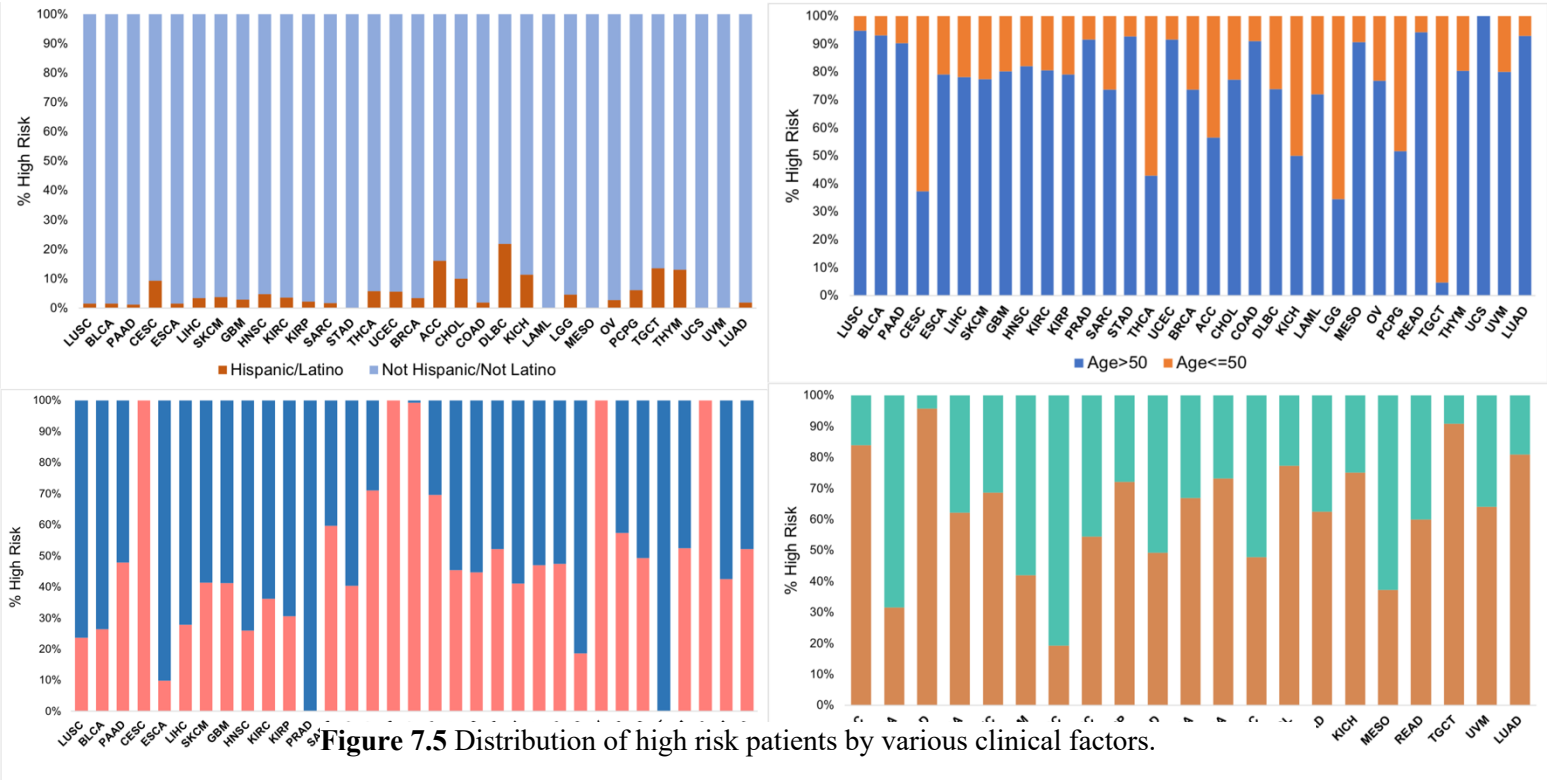


Figure 7.5 Distribution of high risk patients by various clinical factors.

low surviving patients belonging to Stage 3/4. Patients whose age at the time of diagnosis is greater than 50 years is seen to be at a higher survival risk than patients <50 years. This trend is akin to what has already been observed earlier (**Figure 7.2**). Except the cancers which are exclusively related to gender such as gynaecological cancers and breast cancer in the case of females and prostate or testicular cancer in the case of males, the risk is seen to be gender biased in a few cancer types. For example females are at higher risk of death due to thyroid cancer (THCA) and males are at a higher risk of death due to lung cancer (LUSC). These results corroborate the previous epidemiological findings. The figure also shows a huge proportion of high cancer risk patients belinging to the Not Hispanic/latino class, however this result is mostly due to low Hispanic patient data and cannot be relied on. To manage the imbalance between data, chi-square feature selection was implemented and poor features were removed. For example **Table 7.2** shows the selected features for Breast cancer (BRCA) and their corresponding chi square test p-values.

Table 7.2 Chi-square test results for Breast cancer (BRCA) dataset

Feature name	Sub-groups	p-value
gender	Male, Female	<1e-5
menopause_status	Post, Pre, Peri	<1e-5
race	White, Black, Asian	<1e-5
ethnicity	Hispanic, Not hispanic	<1e-5
history_neoadjuvant_treatment	No, Yes	<1e-5
tumor_status	Tumor free, With tumor	<1e-5
method_initial_path_dx	Core needle biopsy Tumor resection Fine needle aspiration biopsy Excisional biopsy Cytology Incisional biopsy	<1e-5
surgical_procedure_first	Modified radical mastectomy Lumpectomy Simple mastectomy	<1e-5
margin_status	Negative, Positive, Close	<1e-5
axillary_staging_method	Axillary lymph node dissection alone Sentinel node biopsy alone Sentinel lymph node biopsy plus axillary dissection No axillary staging	<1e-5
micromet_detection_by_ihc	No, Yes	<1e-5

lymph_nodes_examined	No, Yes	<1e-5
lymph_nodes_examined_count	0-3, 3-9, 9-16, 16-44	2.7e-5
ajcc_tumor_pathologic_pt	T1, T2, T3, T4	<1e-5
ajcc_nodes_pathologic_pn	N0, N1, N2, N3	<1e-5
ajcc_metastasis_pathologic_pm	M0, M1	<1e-5
ajcc_pathologic_tumor_stage	Stage I, Stage II, Stage III, Stage IV	<1e-5
er_status_by_ihc	Positive, Negative	<1e-5
pr_status_by_ihc	Positive, Negative	<1e-5
her2_status_by_ihc	Positive, Negative, Equivocal	<1e-5
histological_subtype	Infiltrating ductal carcinoma Infiltrating lobular carcinoma Mucinous carcinoma Metaplastic carcinoma Medullary carcinoma	<1e-5

7.3.2 Decision trees based risk prediction models

The feature set obtained after the Chi-square selection method was used for building decision tree based classifiers for each cancer. Numerical features were taken as it is where-as categorical features were encoded. Samples with missing feature values were removed. The labels for binary classification was ‘High risk’ or ‘Low risk’ based on the overall survival time and the features

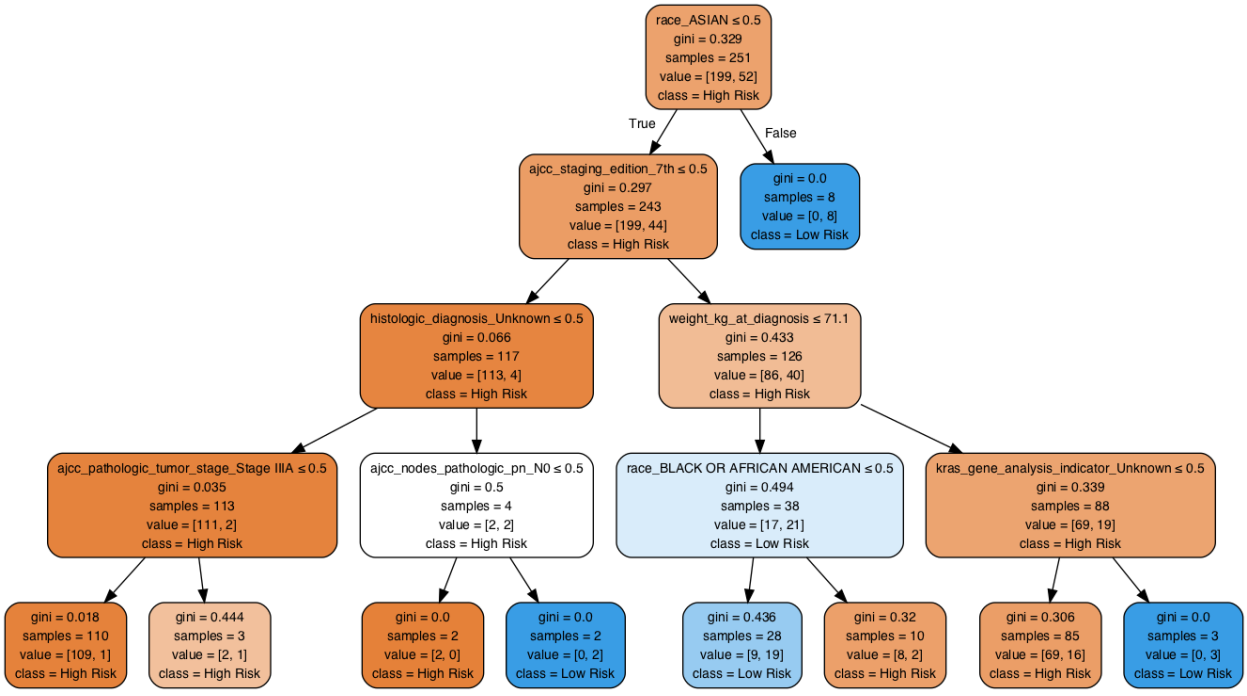


Figure 7.6 Decision tree for risk prediction in Colon Adenocarcinoma (COAD)

were further reduced using a recursive feature elimination technique (RFECV with cv=5). The features were ranked and only the top 10 ranked features were taken for model development. Each decision tree based model was evaluated using a five-fold cross validation scheme. The predicted labels for each cancer were used to stratify risk groups. **Table 7.3** shows the results for corresponding cancer types. As an example **Figure 7.6** shows the decision tree constructed for Colon cancer (COAD). The obtained tree utilized features such as race, weight, ajcc staging and edition and KRAS mutation status for developing the model.

Table 7.3 Decision trees based risk prediction models using clinical factors

Cancer type	Samples	HR	p-value	C
UVM	79	3.47	1.12E-02	0.62
DLBC	47	2.62	4.04E-02	0.59
PRAD	500	2.61	3.70E-02	0.61
KIRP	288	2.32	1.94E-03	0.61
KICH	112	2.09	2.53E-02	0.61
LIHC	299	1.60	3.95E-03	0.56
CHOL	45	1.53	3.55E-02	0.53
ACC	92	1.46	3.13E-02	0.53
PAAD	185	1.41	1.76E-02	0.52
UCEC	445	1.38	3.70E-02	0.55
BRCA	969	1.29	2.27E-02	0.54
CESC	307	1.26	4.04E-02	0.53
MESO	86	1.24	4.35E-02	0.57
LAML	200	1.23	3.39E-02	0.51
ESCA	184	1.22	4.80E-02	0.48
SARC	259	1.20	4.38E-02	0.52
GBM	594	1.18	8.95E-03	0.53
READ	170	1.17	8.36E-02	0.59
LUAD	503	1.16	4.13E-02	0.51
THCA	507	1.13	8.24E-02	0.57
STAD	438	1.12	6.17E-02	0.51
SKCM	345	1.11	5.59E-02	0.52
PCPG	179	1.11	9.13E-02	0.52
OV	584	1.08	5.25E-02	0.50
KIRC	533	1.07	6.72E-02	0.50

HNSC	526	1.04	8.06E-02	0.50
COAD	250	1.03	9.67E-02	0.53
LUSC	491	1.03	8.60E-02	0.51
THYM	123	1.01	9.89E-02	0.48
UCS	57	1.01	9.86E-02	0.52
BLCA	408	1.00	9.90E-02	0.50

*HR: Hazard ratio, C: Concordance Index

In **Table 7.3**, the results for models with an HR value>1.5 and p<0.05 are highlighted in red. The decision tree model for UVM shows the highest risk stratification ability with HR=3.47. The number of features utilized by this model were 14 including features such as M stage, T stage, ethnicity, history of prior disease etc. It should be noted that the risk stratification ability of this model improved in comparison to AJCC Staging in **Table 7.2**. However most of the models performed poorly and weren't able to classify risk groups. Therefore, we implemented another algorithm for model development which is explained in the following section.

7.3.3 Risk matrices and survival prediction

Each subclass of a clinical factor (such as in **Table 7.2**) is attributed a risk probability value. This probability is calculated as explained in materials and methods section. Therefore, the data matrix corresponding to a cancer is converted to a risk matrix. Further, risk vectors are calculated by utilizing different feature columns. A recursive elimination technique is employed to choose the best features. The risk vector corresponding to the best features is the one that significantly stratifies the survival risk groups. Patients with a mean risk probability <0.5 are termed as 'Low-risk' and vice versa. **Table 7.4** shows the result corresponding to this. The model corresponding to Rectal adenocarcinoma (READ) performed the best with an HR of 24.71 and three features including BRAF mutation status, Tumor status and history of past malignancy.

Table 7.4 Risk matrix based risk prediction models using clinical factors

Cancer	Features	HR	p-value	C	%95 CI L	%95 CI U
READ	3	24.71	2.48E-02	0.58	1.50	406.67
THCA	3	22.49	1.83E-06	0.86	6.26	80.76
CHOL	5	21.89	8.32E-06	0.77	5.63	85.03

TGCT	6	13.92	3.55E-02	0.79	1.20	161.92
UCS	4	13.84	4.12E-04	0.69	3.22	59.50
ACC	3	13.04	7.59E-08	0.77	5.11	33.24
PCPG	6	12.91	4.04E-02	0.58	1.12	148.99
THYM	4	9.31	1.10E-02	0.71	1.67	52.01
KIRP	4	8.97	3.56E-09	0.80	4.33	18.59
BRCA	4	8.79	3.07E-02	0.54	1.22	63.11
PRAD	5	8.40	5.80E-03	0.75	1.85	38.09
CESC	2	8.01	4.42E-02	0.51	1.06	60.79
KIRC	6	7.90	1.04E-32	0.75	5.62	11.10
COAD	5	7.62	2.77E-12	0.77	4.31	13.46
LUAD	2	7.61	4.70E-02	0.51	1.03	56.35
LUSC	2	6.37	9.36E-05	0.52	2.52	16.14
KICH	2	6.27	5.00E-03	0.63	1.74	22.58
UCEC	5	5.64	8.59E-05	0.64	2.38	13.37
BLCA	5	5.64	2.69E-09	0.57	3.19	9.97
HNSC	2	5.57	1.12E-24	0.71	4.01	7.74
LIHC	6	4.83	8.00E-03	0.52	1.51	15.49
UVM	3	4.83	1.07E-02	0.69	1.44	16.20
LGG	6	3.61	3.05E-09	0.69	2.36	5.52
STAD	3	2.95	3.76E-03	0.55	1.42	6.12
PAAD	6	2.91	2.50E-05	0.66	1.77	4.79
SKCM	3	2.88	1.46E-09	0.62	2.05	4.06
LAML	4	2.78	5.25E-03	0.55	1.36	5.70
SARC	2	2.72	8.36E-05	0.60	1.65	4.47
ESCA	5	2.47	1.14E-03	0.64	1.43	4.27
GBM	2	2.37	5.27E-05	0.53	1.56	3.60
OV	4	1.66	5.10E-04	0.55	1.25	2.20

*HR: Hazard ratio, C: Concordance Index, CI: Confidence Interval, L: Lower, U: Upper

7.4 Clinical data VS. Molecular data in cancer prognosis

A comparative analysis between the results presented here in Table 7.4 (risk matrix based methods) and bottom ten cancers of Table 6.3 (expression based cancer-specific models) can be presented in the form of the following **Table 7.5**. As observed, the performance is significantly increased in BRCA, LUAD, BLCA, LUSC and HNSC when clinical data is used. For SKCM, the model presented in Chapter 5 is still superior in performance. In all other cancers, both type of models

showed a similar performance. The results, therefore, further emphasize the importance of using clinical data as opposed to more sophisticated omics based approached.

Table 7.5 The table shows the comparsion between clinical data based models, HR(CL) and expression data based models from Chapter 6, HR (EXP).

Cancer	HR (EXP)	p-value	C	HR (CL)	p-value	C
SKCM	1.99	$2.18x10^{-5}$	0.59	2.88	1.46E-09	0.62
GBM	2.07	$3.73x10^{-4}$	0.61	2.37	5.27E-05	0.53
OV	2.19	$1.38x10^{-6}$	0.61	1.66	5.10E-04	0.55
LUSC	2.21	$1.26x10^{-6}$	0.61	6.37	9.36E-05	0.52
HNSC	2.36	$9.24x10^{-8}$	0.62	5.57	1.12E-24	0.71
LUAD	2.76	$6.94x10^{-8}$	0.63	7.61	4.70E-02	0.51
SARC	2.81	$1.32x10^{-5}$	0.67	2.72	8.36E-05	0.6
STAD	3.35	$2.78x10^{-7}$	0.64	2.95	3.76E-03	0.55
BLCA	3.41	$6.35x10^{-10}$	0.66	5.64	2.69E-09	0.57
BRCA	3.45	$2.36x10^{-9}$	0.67	8.79	3.07E-02	0.54

*HR: Hazard Ratio, C: Concordance Index, EXP: Expression based, CL: Clinical data based

7.5 Conclusion and summary

Risk evaluation is a crucial step in cancer management. A careful prognosis is often required for strategic planning of therapeutic intervention. This has lead to development of several prognostic methods and identification of biomarkers. However, modern oncology research seems to be biased towards omics based techniques and consistently ignores the contribution and role of other intrinsic and extrinsic risk factors. Past studies have revealed important roles of various factors in the development of cancer such as tobacco smoking in lung cancer, radiation exposure in thyroid cancer etc. On on hand, some factors can be modified/controlled as prevention tactics for cancer, on other hand some of these can be exploited for risk evaluation in cancer patients. In the current study, we examined a plethora of 'clinical factors' obtained from monitoring of a large number of cancer patients. The prognostic strength of these factors was probed by two approaches, first being the machine learning based classification of patients into risk groups. Since, the majority of the considered factors were categorical, decision trees which handle both numerical and categorical features, were used to build ML models. Second, approach involved construction of a 'risk matrix'

wherein each entry replaces a clinical “characteristic” with a probability value representing the likelihood of death risk. A few prognostic models based on decision tree based method was seen to perform better than the conventional staging, whereas the risk matrix based approach proved to be superior and provided better stratification models. These models employ minimal number of features and also have an anhanced prognostic potential as compared to conventional staging. Overall this study highlights the strength of clinical factors in cancer prognosis, and motivates further research into the untapping the potential of such factors for advancement in cancer care and management.

8

Summary and Conclusion

Cancer is the second leading cause of mortality globally (Siegel *et al.*, 2020). For treatment and management of cancer patients, a crude pipeline that is generally followed by oncologists worldwide consists of the following broad steps (a) Screening: Early physical examinations or lab tests on suspicion of cancer and also various tumour biomarker/marker tests (b) Diagnosis: which includes imaging tests such as CT scans, PET, MRI, X-rays etc., invasive procedures such as biopsy and endoscopy and/or blood or genetic tests for cancer biomarkers (c) Risk evaluation: primarily includes cancer staging for evaluation of severity and prognosis and (d) Therapeutic decision making involving surgical resection if cancer is localized and/or therapies such as chemotherapy , radiotherapy , immunotherapy etc. if cancer has metastasized to other tissues/body parts. Since, therapeutic intervention in majority of the cases follows the cancer staging system, it is considered to be the most important step in clinical management of cancer patients. For most of the cancers, the guidelines for staging is provided by American Joint Committee on Cancer (AJCC) and the Union for International Cancer Control (UICC). This is commonly known as the TNM staging system. As discussed in **chapter 1**, in the TNM system, the overall stage is determined after the cancer is assigned a letter or number to describe the tumor (T), node (N), and metastasis (M) categories wherein, T describes the size and location of original (primary) tumor, N tells whether the cancer has spread to the nearby lymph nodes and M tells whether the cancer has spread (metastasized) to distant parts of the body. The staging system many a times also includes information about other features such as levels of some molecular markers (PSA in Prostate cancer), age of the patient (Thyroid cancer), presence/absence of specific proteins (ER, PR or HER in Breast cancer), tumor location (esophageal cancer) etc. Although, not all cancers are staged using the TNM system, for example Staging systems other than the TNM system are often used for Hodgkin and non-Hodgkin lymphomas, as well as for some other cancers. The International Federation of Gynecologists and Obstetricians (FIGO) has a staging system for cancers of the female reproductive organs. However, the TNM stages closely match the FIGO stages, which makes it fairly easy to convert stages between these 2 systems. Once the values for T, N, and M (and any other factors that affect stage) have been determined, they are combined to assign an overall stage. For most cancers, the stage is a Roman numeral from I (1) to IV (4). Stage I cancers

are less advanced and often have a better prognosis. Higher stage cancers typically have spread farther, so they might require more intense or different treatment. However, due to complexities associated with cancer, modern oncologists are always learning more about cancer growth and progression, and best treatment strategies. Over time, some of these findings are added to the staging systems for different types of cancer, which helps make them more accurate and valuable to both doctors and patients. Notable examples include addition of Gleason Score (Chen and Zhou, 2016) in prostate cancer staging; ER, PR, and HER2 in breast cancer staging; and LDH in melanoma (Gershenwald *et al.*, 2017) staging in the AJCC 8th edition staging (Amin *et al.*, 2017). The inclusion of HER2 status was a result of a new Neo-Bioscore staging system developed by researchers at MD Anderson Cancer Center, thereby allowing more precise prognostic stratification of all breast cancer subtypes (Mittendorf *et al.*, 2016).

A huge challenge in cancer biomarker development is the heterogeneity associated with cancer since each cancer is composed of varied phenotypes and often responds differently to same therapeutic intervention. This heterogeneity arises due to the aberrant behavior in the cells of an individual cancer type. To tackle this, modern oncologists are continuously putting efforts to gain a detailed knowledge of the cellular mechanisms that drive cancer. It is now believed that biomarker development utilizing the genomic and proteomic information is a superior way of carefully addressing the issue of cancer heterogeneity. Thus, identification of novel biomarkers, now-a-days, largely relies on the "omics" technologies. The earlier notion of a single biomarker has now been replaced with multi-panel biomarkers or signatures consisting of genes or proteins thereby revealing the vital fingerprint correlated with a given cancer.

Genomics has been used widely for the detection and recognition of biomarkers. The availability of genome sequencing technologies and microarray expression methods [29] provide a reliable and minimally invasive feature extraction system. This helps researchers to go another step forward, developing and producing a biological drug with a deeper knowledge of pharmacogenomics, thereby allowing biomarkers to investigate the effects of genetic variation, creating novel strategies for personally treating patients. In this study, we focused on a prominent cellular mechanism, apoptosis, which has a strong and proven foundation in cancer growth and development, as presented in **chapter 1 and 2**. In **chapter 4** of this thesis, we show that certain genes belonging to the apoptotic pathway are correlated with patient survival in Thyroid cancer

(papillary thyroid carcinoma). The elevation and suppression of mRNA levels of these genes may be responsible in an aggressive or a mild phenotype of cancer thereby affecting patient outcome. The proposed signature in a further analysis was seen to perform better than AJCC staging for risk stratification purposes. A comprehensive evaluation of other clinical factors motivated the addition of patient age to this signature, thereby resulting in a genomic-clinical hybrid panel. The identified genes also show a differential behavior amongst normal and cancerous tissue implying their power to distinguish between people with cancer and without cancer. To guide this study in the direction of personalized medicine, candidate drug molecules were identified which could potentially modulate the expression levels of both adverse and beneficial genes and potentially reduce the severity of the disease.

Chapter 6 of this thesis also utilized the genomic data of apoptotic pathway genes in order to identify universal gene signatures which hold prognostic value across different cancer types. This is in contrast to the typical process of development of cancer-specific biomarkers. The study centered at identification of prognostic biomarker apoptotic genes across different cancer types and devised a 11 gene panel that is applicable across 27 cancer types. Though performance of the panel is seen to vary amongst cancer types, a significant stratification is achieved in all the cases. In addition to this, the analysis presented in the chapter also offered a novel strategy of cross cancer biomarker development and sheds light on a novel gene signature which can be used in both brain cancer and kidney cancer patients. Apart from its prognostic relevance, the underlying nature of the genes could also motivate development of common therapeutic strategies in cells with different types of origin (glial cells in nervous system vs. tissue lining cells in kidneys). Further, the study also put forward cancer-specific risk prediction models based on expression levels of apoptotic genes.

Gene expression profiling is a very reliable technique for classification of Cancer and prognostication, though, in the form of signalling networks, the function of these genes depends on their translation into functional proteins. This understanding is achieved by studying the proteins through the application of proteomics. Proteomics is based on the analysis of determination of levels of translated proteins in a given specimen, tissue or organism. Since, the fundamental protein families regulating the apoptotic pathway along with their functions is largely understood (**Chapter 2**), it is expected that an in-depth analysis of the proteomic profiles of

different tissue sampled collected from cancer patients would improve our knowledge of tumour pathogenesis, prognosis, and identification of therapeutic targets. To this endeavor, **Chapter 3** incorporated a proteomic dataset with expression profile of Bcl2 family proteins in the context of colo-rectal cancer. The study analyzed different protein expression based models and developed a novel protein signature for predicting (Folflox and Xelox) therapy responders and non-responders in Stage III CRC patients. The proposed signature assigns each patient with a 'risk score' based on the expression value of 5 pro- and anti-apoptotic proteins. A greater score is indicative of failure of therapy and higher mortality risk for the given patient. This study illuminates the synergistic role of the proteins in conferring therapeutic resistance in CRC and vital role of apoptosis. As a practical applicability of the proposed model, the in-house web-server 'CRCRpred' (https://webs.iiitd.edu.in/raghava/crcrpred) can be further exploited by both clinicians and patients. This resource can be beneficial in therapy planning and personalized treatment.

The aforesaid "omics" technologies and subsequent data has led to the development of an extensive variety of cancer biomarkers, for cancer risk assessment purposes. This also includes the biomarkers/models developed in the current study. However, at the same time, much of these newer findings often makes the staging systems more complex than they were in the past, which can make it harder for people to understand them. Therefore, despite their excellent performance in the cohort studies, majority of the biomarkers haven't yet been added to the staging systems. For that reason, our current study also explores the roles of various 'clinical factors' which collectively include pathological features, demographic features, lifestyle related features, anatomic features, blood protein level status (such as ER) etc. in predicting survival outcome of cancer patients. **Chapter 4** of the thesis thoroughly examines the prognostic strength of genomic data corresponding to cancer-associated pathways as well as clinical factors in Melanoma patients. Multiple gene expression-based risk prediction models are developed and evaluated in comparison with clinical factors. Models were also constructed based on combination of best genomic features and clinical features. However, the model which had only clinical factors performed superior to all the other models. This study therefore highlights the importance of clinical factors in risk assessment. It indicates how a schematic integration of existing clinical features in the staging process can be more efficient. It also hints that, while, the omics-based biomarkers can be alluring due to their innate biological association, clinical factors should not be undermined. Based on this

pretext, the analysis presented in **Chapter 7** of the thesis takes on a pan-cancer approach of developing risk prediction models by employing clinical factors only. The clinical factors herein are inclusive of a wide range of features spanning from intrinsic or heritable factors, various extrinsic risk factors, anatomical features and surgical methods or therapy procedures employed. The goal of the study was to develop risk prediction models which are easy to implement and comprehend. The models were assessed against the staging schemes in different cancer for their efficacy in risk evaluation in lieu of current standards.

Overall, the work presented in this thesis proposes several novel prognostic biomarkers and methods for better risk evaluation in cancer patients. On one hand, the pipeline used in the study exploited a crucial cellular mechanism by utilizing recent "omics" based data and modelling techniques. On the other hand, various clinical features were evaluated both individually and as combination to suggested biomarker genes/proteins in lieu of patient survival. The comprehensive analysis of apoptosis molecules shed light on the risk prediction ability of the expression data in various cancers such as protein expression as therapeutic predictive marker in Colon and Rectal cancer (Chapter 3) and gene expression in Thyroid cancer (Chapter 4). However, some exceptional cases were also observed such as Melanoma where apoptosis expression based prognostic markers failed to stratify risk groups efficiently. This is clearly observed in 'cancer-specific' models presented in Chapter 6 where Model for SKCM has the lowest performance. Thus, this specific case was exclusively addressed in Chapter 5, wherein other pathways and clinical features were studied. This resulted in a solely clinical features based risk model and trumped various other expression based models. The enhancement due to clinical feature addition is also obvious in hybrid approaches implemented in Chapter 3 and 4. Subsequently, a thorough analysis of clinical features was performed in Chapter 7. Following this, a comparison between expression based and clinical data based models was established.

This work addresses various aspects of molecular and clinical data in prognosis of cancer patients, however, it will be naïve to say that the approaches taken here are complete/accurate. Ideally, a more holistic way of model development is required, which has to involve several other factors such as epigenomics, metabolomics, single-cell studies etc. owing to the heterogenous nature of the disease. Herein, although, we utilized a reductionist approach, due to several limitations pertaining to time, resources and computational power, the inclusion of clinical data can be

considered as a step towards holistic understanding. To conclude, the work presented here, in its current form albeit after thorough clinical validations, can be beneficial for designing better treatment strategies and thereby help in progress of cancer research.

9

Appendix A

APPENDIX A

Table A1 The list of clinical features corresponding to each cancer-type (Chapter-7), before feature selection was implemented. The feature annotation follows the TCGA annotation for clinical data.

Cancer	Features
ACC	gender, race, ethnicity, history_other_malignancy, history_neoadjuvant_treatment, tumor_status, residual_tumor, history_adrenal_hormone_excess, age_at_diagnosis, cytoplasm_presence_less_than_equal_25_percent, clinical_M, pathologic_T
BLCA	gender, height_cm_at_diagnosis, weight_kg_at_diagnosis, race, ethnicity, history_other_malignancy, history_neoadjuvant_treatment, noninvasive_bladder_history, tumor_status, occupation_current, tobacco_smoking_history_indicator, radiation_treatment_adjuvant, tobacco_smoking_pack_years_smoked, pharmaceutical_tx_adjuvant, histologic_subtype, age_at_diagnosis, ajcc_staging_edition, ajcc_tumor_pathologic_pt, lymphovascular_invasion, ajcc_nodes_pathologic_pn, lymph_nodes_examined, lymph_nodes_examined_count, lymph_nodes_examined_he_count, ajcc_metastasis_pathologic_pm, ajcc_pathologic_tumor_stage, incidental_prostate_cancer_indicator, new_tumor_event_dx_indicator, anatomic_neoplasm_subdivision, histological_type, neoplasm_histologic_grade
BRCA	gender, menopause_status, race, ethnicity, history_neoadjuvant_treatment, tumor_status, age_at_diagnosis, method_initial_path_dx, surgical_procedure_first, margin_status, axillary_staging_method, micromet_detection_by_ihc, lymph_nodes_examined, lymph_nodes_examined_count, ajcc_tumor_pathologic_pt, ajcc_nodes_pathologic_pn, ajcc_metastasis_pathologic_pm, ajcc_pathologic_tumor_stage, er_status_by_ihc, pr_status_by_ihc, her2_status_by_ihc, histological_type

CESC	gender, menopause_status, height_cm_at_diagnosis, weight_kg_at_diagnosis, race, history_hormonal_contraceptives_use, pregnancies_count_total, pregnancies_count_live_birth, history_neoadjuvant_treatment, tumor_status, pregnant_at_diagnosis, ecog_score, age_at_diagnosis, history_other_malignancy, histologic_diagnosis, keratinization_squamous_cell, tumor_grade, ajcc_nodes_pathologic_pn, hysterectomy_type, lymph_nodes_examined, lymph_nodes_examined_count, ajcc_tumor_pathologic_pt, ajcc_metastasis_pathologic_pm, ajcc_staging_edition, radiation_treatment_adjuvant.1, pharmaceutical_tx_adjuvant.1, clinical_stage
CHOL	gender, height_cm_at_diagnosis, weight_kg_at_diagnosis, race, ethnicity, history_other_malignancy, history_neoadjuvant_treatment, tumor_status, family_history_cancer_indicator, family_history_cancer_relationship, history_hepato_carcinoma_risk_factors, radiation_treatment_adjuvant, pharmaceutical_tx_adjuvant, ablation_embolization_tx_adjuvant, histologic_diagnosis, definitive_surgical_procedure, tumor_grade, residual_tumor, ajcc_tumor_pathologic_pt, ajcc_nodes_pathologic_pn, ajcc_metastasis_pathologic_pm, ajcc_pathologic_tumor_stage, vascular_invasion, perineural_invasion, child_pugh_classification, ca_19_9_level, alpha_fetoprotien_at_procurement, platelet_count_preresection, prothrombin_time_INR_at_procurement, serum_albumin_preresection, bilirubin_total, creatinine_level_preresection, ishak_fibrosis_score, ecog_score, new_tumor_event_dx_indicator, age_at_diagnosis
COAD	histologic_diagnosis, gender, race, ethnicity, history_other_malignancy, history_neoadjuvant_treatment, ajcc_staging_edition, ajcc_tumor_pathologic_pt, ajcc_nodes_pathologic_pn, ajcc_metastasis_pathologic_pm, ajcc_pathologic_tumor_stage, residual_tumor, tumor_status, cea_level_pretreatment, vascular_invasion_indicator, lymphovascular_invasion_indicator, kras_gene_analysis_indicator, braf_gene_analysis_indicator, history_other_malignancy.1, history_colon_polyps, weight_kg_at_diagnosis, height_cm_at_diagnosis, family_history_colorectal_cancer, age_at_diagnosis, anatomic_neoplasm_subdivision

DLBC	histologic_diagnosis, history_other_malignancy, history_neoadjuvant_treatment, gender, race, ethnicity, weight_kg_at_diagnosis, height_cm_at_diagnosis, age_at_diagnosis, clinical_stage
ESCA	gender, height_cm_at_diagnosis, weight_kg_at_diagnosis, race, ethnicity, history_other_malignancy, tumor_status, esophageal_tumor_location_centered, esophageal_tumor_location_involved, histologic_diagnosis, tumor_grade, age_at_diagnosis, ajcc_staging_edition, ajcc_tumor_pathologic_pt, ajcc_nodes_pathologic_pn, ajcc_metastasis_pathologic_pm, ajcc_pathologic_tumor_stage, residual_tumor
GBM	gender, race, ethnicity, history_neoadjuvant_treatment, tumor_status, karnofsky_score, age_at_diagnosis, histological_type
HNSC	anatomic_organ_subdivision, laterality, gender, race, history_other_malignancy, history_neoadjuvant_treatment, lymph_node_neck_dissection_indicator, lymph_nodes_examined, lymph_nodes_examined_count, lymph_nodes_examined_he_count, margin_status, tumor_status, ajcc_staging_edition, ajcc_tumor_pathologic_pt, ajcc_nodes_pathologic_pn, ajcc_pathologic_tumor_stage, extracapsular_spread_pathologic, tumor_grade, lymphovascular_invasion, perineural_invasion, tobacco_smoking_history_indicator, alcohol_history_documented, age_at_diagnosis, clinical_M, clinical_N, clinical_T, clinical_stage, tissue_source_site
KICH	histologic_diagnosis, sarcomatoid_features, laterality, gender, race, ethnicity, history_other_malignancy, history_neoadjuvant_treatment, ajcc_tumor_pathologic_pt, ajcc_nodes_pathologic_pn, ajcc_metastasis_clinical_cm, ajcc_pathologic_tumor_stage, age_at_diagnosis
KIRC	histologic_diagnosis, tumor_grade, laterality, gender, race, ethnicity, history_other_malignancy, history_neoadjuvant_treatment, ajcc_tumor_pathologic_pt, ajcc_nodes_pathologic_pn, ajcc_metastasis_pathologic_pm, ajcc_pathologic_tumor_stage, tumor_status, serum_calcium_level, hemoglobin_level, platelet_count, white_cell_count, age_at_diagnosis

KIRP	histologic_diagnosis, tumor_type, laterality, gender, race, ethnicity, height_cm_at_diagnosis, weight_kg_at_diagnosis, history_other_malignancy, history_neoadjuvant_treatment, age_at_diagnosis, lymph_nodes_examined, ajcc_staging_edition, ajcc_tumor_clinical_ct, ajcc_nodes_clinical_cn, ajcc_metastasis_clinical_cm, ajcc_clinical_tumor_stage, ajcc_tumor_pathologic_pt, ajcc_nodes_pathologic_pn, ajcc_metastasis_pathologic_pm, ajcc_pathologic_tumor_stage, tumor_status
LAML	gender, race, ethnicity, history_other_malignancy, history_hematologic_disorder, history_neoadjuvant_treatment, history_neoadjuvant_hydroxyurea_tx, history_exposure_leukemogenic_agents, cells_used_for_analysis_source, age_at_diagnosis, percent_blasts_peripheral_blood, fab_category, cyto_and_immuno_test_performed, cyto_and_immuno_test_percentage, percent_cellularity, wbc_24hr_of_banking, hemoglobin_24hr_of_banking, platelet_count_preresection, blast_count, promyelocytes_count, segs_24hr_of_banking, basophils_count, abnormal_lymphocyte_percent, promonocytes_24hr_of_banking, fish_abnormality_detected, test_performed_indicator, fish_performed_outcome, molecular_studies_others_performed, molecular_abnormality_results, molecular_abnormality_percent, atra_exposure, informed_consent_verified
LIHC	gender, height_cm_at_diagnosis, weight_kg_at_diagnosis, race, history_neoadjuvant_treatment, tumor_status, family_history_cancer_indicator, history_hepato_carcinoma_risk_factors, radiation_treatment_adjuvant, pharmaceutical_tx_adjuvant, ablation_embolization_tx_adjuvant, histologic_diagnosis, definitive_surgical_procedure, tumor_grade, residual_tumor, ajcc_staging_edition, ajcc_tumor_pathologic_pt, ajcc_nodes_pathologic_pn, ajcc_metastasis_pathologic_pm, ajcc_pathologic_tumor_stage, vascular_invasion, child_pugh_classification, alpha_fetoprotien_at_procurement, alpha_fetoprotien_norm_range_lower, platelet_count_preresection, platelet_norm_range_lower, prothrombin_time_INR_at_procurement, serum_albumin_preresection, bilirubin_total_norm_range_upper, bilirubin_total_norm_range_lower, age_at_diagnosis

LUAD	histologic_diagnosis, gender, submitted_tumor_site, race, ethnicity, history_other_malignancy, anatomic_organ_subdivision, histologic_diagnosis.1, residual_tumor, ajcc_staging_edition, ajcc_tumor_pathologic_pt, ajcc_nodes_pathologic_pn, ajcc_metastasis_pathologic_pm, pulmonary_function_test_indicator, kras_gene_analysis_indicator, egfr_mutation_status, tobacco_smoking_history_indicator, history_neoadjuvant_treatment, tumor_status, age_at_diagnosis
LUSC	histologic_diagnosis, gender, race, ethnicity, history_other_malignancy, anatomic_organ_subdivision, histologic_diagnosis.1, residual_tumor, ajcc_staging_edition, ajcc_tumor_pathologic_pt, ajcc_nodes_pathologic_pn, ajcc_metastasis_pathologic_pm, ajcc_pathologic_tumor_stage, egfr_mutation_status, eml4_alk_translocation_status, history_neoadjuvant_treatment, tumor_status, age_at_diagnosis
MESO	gender, race, ethnicity, history_other_malignancy, history_neoadjuvant_treatment, pleurodesis_performed_prior, tumor_status, history_asbestos_exposure, primary_occupation, occupation_primary, radiation_treatment_adjuvant, pharmaceutical_tx_adjuvant, laterality, histologic_diagnosis, ajcc_tumor_pathologic_pt, ajcc_nodes_pathologic_pn, ajcc_metastasis_pathologic_pm, ajcc_pathologic_tumor_stage, residual_tumor, creatinine_prior_tx, mesothelioma_detection_method, age_at_diagnosis
OV	gender, race, ethnicity, history_neoadjuvant_treatment, tumor_status, tumor_grade, residual_disease_largest_nodule, age_at_diagnosis, anatomic_neoplasm_subdivision, clinical_stage, histological_type
PAAD	invasive_adenocarcinoma_indicator, histologic_diagnosis, tumor_sample_type, gender, race, ethnicity, history_other_malignancy, history_neoadjuvant_treatment, tumor_grade, grade_tier_system, tumor_resected_max_dimension, residual_tumor, ajcc_staging_edition, ajcc_tumor_pathologic_pt, ajcc_nodes_pathologic_pn, ajcc_metastasis_pathologic_pm, ajcc_pathologic_tumor_stage, tumor_status, diabetes_diagnosis_indicator, history_chronic_pancreatitis, family_history_cancer_indicator, radiation_treatment_adjuvant, age_at_diagnosis, anatomic_neoplasm_subdivision

PCPG	gender, race, ethnicity, history_other_malignancy, history_pheo_or_para_include_benign, history_neoadjuvant_treatment, tumor_status, laterality, histologic_diagnosis, age_at_diagnosis
PRAD	histologic_diagnosis, zone_of_origin, gleason_pattern_primary, gleason_pattern_secondary, gleason_score, laterality, tumor_level, gender, history_other_malignancy, history_neoadjuvant_treatment, ct_scan_ab_pelvis_indicator, mri_at_diagnosis, lymph_nodes_examined, lymph_nodes_examined_count, lymph_nodes_examined_he_count, residual_tumor, tumor_status, biochemical_recurrence_indicator, radiation_treatment_adjuvant, new_tumor_event_dx_indicator, age_at_diagnosis, clinical_M, clinical_T, pathologic_T, targeted_molecular_therapy
READ	histologic_diagnosis, gender, race, history_other_malignancy, history_neoadjuvant_treatment, ajcc_tumor_pathologic_pt, ajcc_nodes_pathologic_pn, ajcc_metastasis_pathologic_pm, ajcc_pathologic_tumor_stage, residual_tumor, tumor_status, braf_gene_analysis_indicator, kras_gene_analysis_indicator, history_other_malignancy.1, family_history_colorectal_cancer, age_at_diagnosis, anatomic_neoplasm_subdivision
SARC	gender, race, ethnicity, history_other_malignancy, history_neoadjuvant_treatment, tumor_status, histologic_diagnosis, age_at_diagnosis, margin_status, residual_tumor, tumor_total_necrosis, disease_multifocal_indicator, locoregional_recurrence_indicator, metastatic_disease_confirmed, nte_lesion_radiologic_length
SKCM	gender, weight_kg_at_diagnosis, race, history_other_malignancy, history_neoadjuvant_treatment, tumor_status, breslow_thickness_at_diagnosis, clark_level_at_diagnosis, primary_melanoma_tumor_ulceration, age_at_diagnosis, ajcc_staging_edition, ajcc_tumor_pathologic_pt, ajcc_nodes_pathologic_pn, ajcc_metastasis_pathologic_pm, ajcc_pathologic_tumor_stage
STAD	histologic_diagnosis, tumor_grade, gender, race, ethnicity, history_other_malignancy, history_neoadjuvant_treatment, residual_tumor, ajcc_staging_edition, ajcc_tumor_pathologic_pt, ajcc_nodes_pathologic_pn, ajcc_metastasis_pathologic_pm, ajcc_pathologic_tumor_stage, family_history_of_stomach_cancer, age_at_diagnosis, anatomic_neoplasm_subdivision

TGCT	race, ethnicity, history_other_malignancy, history_of_undescended_testis, history_hypospadias, history_fertility, family_history_testicular_cancer, family_history_other_cancer, history_neoadjuvant_treatment, tumor_status, laterality, testis_tumor_macroextent, histologic_diagnosis, histologic_diagnosis_percent, intratubular_germ_cell_neoplasm, ajcc_staging_edition, ajcc_tumor_clinical_ct, ajcc_nodes_clinical_cn, ajcc_metastasis_clinical_cm, ajcc_clinical_tumor_stage, ajcc_tumor_pathologic_pt, ajcc_nodes_pathologic_pn, ajcc_metastasis_pathologic_pm, ajcc_pathologic_tumor_stage, serum_markers, pre_orchi_hcg, first_treatment_success, age_at_diagnosis, gender
THCA	gender, race, ethnicity, history_other_malignancy, history_neoadjuvant_treatment, tumor_status, history_thyroid_disease, history_radiation_exposure, histologic_diagnosis, laterality, tumor_focality, tumor_size_width, tumor_size_width.1, tumor_size_width.2, age_at_diagnosis, lymph_nodes_preop_imaging, lymph_nodes_preop_imaging_type, lymph_nodes_examined, lymph_nodes_examined_count, lymph_nodes_examined_he_count, extrathyroidal_extension, residual_tumor, ajcc_staging_edition, ajcc_tumor_pathologic_pt, ajcc_nodes_pathologic_pn, ajcc_metastasis_pathologic_pm, ajcc_pathologic_tumor_stage, genotypic_analysis_detected
THYM	gender, height_cm_at_diagnosis, weight_kg_at_diagnosis, race, ethnicity, history_other_malignancy, history_neoadjuvant_treatment, tumor_status, radiation_treatment_adjuvant, pharmaceutical_tx_adjuvant, ablation_embolization_tx_adjuvant, method_initial_path_dx, masaoka_stage, history_myasthenia_gravis, new_tumor_event_dx_indicator, age_at_diagnosis
UCEC	gender, menopause_status, height_cm_at_diagnosis, weight_kg_at_diagnosis, race, ethnicity, history_neoadjuvant_treatment, tumor_status, histologic_diagnosis, age_at_diagnosis, method_initial_path_dx, surgical_approach_at_diagnosis, peritoneal_washing, tumor_invasion_percent, residual_tumor, lymph_nodes_pelvic_examined_count, lymph_nodes_pelvic_pos__by_he, clinical_stage, neoplasm_histologic_grade

UCS	gender, menopause_status, history_menopausal_hormone_therapy, history_tamoxifen_use, hypertension_diagnosis, diabetes_diagnosis_indicator, pregnancies_full_term_count, history_colorectal_cancer, height_cm_at_diagnosis, weight_kg_at_diagnosis, race, ethnicity, history_other_malignancy, history_neoadjuvant_treatment, tumor_status, vital_status, treatment_outcome_first_course, radiation_treatment_adjuvant, pharmaceutical_tx_adjuvant, surgical_approach_at_diagnosis, peritoneal_washing, tumor_invasion_percent, lymph_nodes_pelvic_examined_count, lymph_nodes_aortic_examined_count, new_tumor_event_dx_indicator, age_at_diagnosis, anatomic_neoplasm_subdivision, clinical_stage, residual_tumor
UVM	gender, height_cm_at_diagnosis, weight_kg_at_diagnosis, race, ethnicity, history_other_malignancy, history_neoadjuvant_treatment, tumor_status, histologic_diagnosis.1, tumor_thickness, tumor_thickness_measurement, ajcc_tumor_clinical_ct, ajcc_metastasis_clinical_cm, ajcc_clinical_tumor_stage, ajcc_tumor_pathologic_pt, ajcc_nodes_pathologic_pn, ajcc_metastasis_pathologic_pm, ajcc_pathologic_tumor_stage, age_at_diagnosis

References

Aguirre-Gamboa,R. *et al.* (2013) SurvExpress: an online biomarker validation tool and database for cancer gene expression data using survival analysis. *PLoS One*, **8**, e74250.

Amin,M.B. *et al.* (2017) The Eighth Edition AJCC Cancer Staging Manual: Continuing to build a bridge from a population-based to a more 'personalized' approach to cancer staging. *CA. Cancer J. Clin.*, **67**, 93–99.

Arora,C. *et al.* (2020) Risk prediction in cutaneous melanoma patients from their clinico-pathological features: superiority of clinical data over gene expression data. *Heliyon*, **6**.

Ashford,N.A. *et al.* (2015) Cancer risk: role of environment. *Science*, **347**, 727.

Bai,Z. *et al.* (2011) Proteomics-based identification of a group of apoptosis-related proteins and biomarkers in gastric cancer. *Int. J. Oncol.*, **38**, 375–383.

Bauer,J.H. and Helfand,S.L. (2006) New tricks of an old molecule: lifespan regulation by p53. *Aging Cell*, **5**, 437–440.

Beyer,H.L. *et al.* (2007) MESOMARK: a potential test for malignant pleural mesothelioma. *Clin. Chem.*, **53**, 666–672.

Bhalla,S. *et al.* (2020) Expression based biomarkers and models to classify early and late-stage samples of Papillary Thyroid Carcinoma. *PLoS One*, **15**, e0231629.

Bian,J. *et al.* (2020) Identification of a five-gene signature for predicting the progression and prognosis of stage I endometrial carcinoma. *Oncol. Lett.*, **20**, 2396–2410.

Boffetta,P. and Hashibe,M. (2006) Alcohol and cancer. *Lancet. Oncol.*, **7**, 149–156.

Borcherding,N. *et al.* (2018) TRGAted: A web tool for survival analysis using protein data in the Cancer Genome Atlas. *F1000Research*, **7**, 1235.

Braillon,A. (2018) Alcohol: Cardiovascular Disease and Cancer. *J. Am. Coll. Cardiol.*, **71**, 582–583.

Brandao,M. *et al.* (2019) Mammaprint: a comprehensive review. *Future Oncol.*, **15**, 207–224.

Bray,F. and Møller,B. (2006) Predicting the future burden of cancer. *Nat. Rev. Cancer*, **6**, 63–74.

Bredesen,D.E. *et al.* (2006) Cell death in the nervous system. *Nature*, **443**, 796–802.

Brown,J.C. *et al.* (2012) Cancer, physical activity, and exercise. *Compr. Physiol.*, **2**, 2775–2809.

Brozyna,A.A. *et al.* (2017) TRPM1 (melastatin) expression is an independent predictor of overall survival in clinical AJCC stage I and II melanoma patients. *J Cutan Pathol*, **44**, 328–337.

Burstein,H.J. *et al.* (2001) Clinical activity of trastuzumab and vinorelbine in women with HER2-overexpressing metastatic breast cancer. *J. Clin. Oncol.*, **19**, 2722–2730.

Chang,Y.-S. *et al.* (2018) Detection of Molecular Alterations in Taiwanese Patients with Medullary Thyroid Cancer Using Whole-Exome Sequencing. *Endocr. Pathol.*, **29**, 324–331.

Charles,E.M. and Rehm,M. (2014) Key regulators of apoptosis execution as biomarker candidates in melanoma. *Mol. Cell. Oncol.*, **1**, e964037.

Chen,N. and Zhou,Q. (2016) The evolving Gleason grading system. *Chin. J. Cancer Res.*, **28**, 58–64.

Chen,W. *et al.* (2020) Transcriptomic analysis reveals that heat shock protein 90α is a potential diagnostic and prognostic biomarker for cancer. *Eur. J. cancer Prev. Off. J. Eur. Cancer Prev. Organ.*, **29**, 357–364.

Cho,S.Y. *et al.* (2017) A novel combination treatment targeting BCL-XL and MCL1 for KRAS/BRAF-mutated and BCL2L1-amplified colorectal cancers. *Mol. Cancer Ther.*, **16**, 2178–2190.

Christoforidou,E.P. *et al.* (2013) Bladder cancer and arsenic through drinking water: a systematic review of epidemiologic evidence. *J. Environ. Sci. Health. A. Tox. Hazard. Subst. Environ. Eng.*, **48**, 1764–1775.

Cook,R.W. *et al.* (2018) Analytic validity of DecisionDx-Melanoma, a gene expression profile test for determining metastatic risk in melanoma patients. *Diagn Pathol*, **13**, 13.

Correa,H. (2016) Li-Fraumeni Syndrome. *J. Pediatr. Genet.*, **5**, 84–88.

Csuka,O. *et al.* (1997) Predictive value of p53, Bcl2 and bax in the radiotherapy of head and neck cancer. *Pathol. Oncol. Res.*, **3**, 204–210.

Deichmann,M. *et al.* (2004) Diagnosing melanoma patients entering American Joint Committee on Cancer stage IV, C-reactive protein in serum is superior to lactate dehydrogenase. *Br J Cancer*, **91**, 699–702.

Ding,L. *et al.* (2020) KIF15 facilitates gastric cancer via enhancing proliferation, inhibiting apoptosis, and predict poor prognosis. *Cancer Cell Int.*, **20**, 125.

Dong,T. *et al.* (2017) WNT10A/betacatenin pathway in tumorigenesis of papillary thyroid carcinoma. *Oncol. Rep.*, **38**, 1287–1294.

Donington,J.S. and Colson,Y.L. (2011) Sex and gender differences in non-small cell lung cancer. *Semin. Thorac. Cardiovasc. Surg.*, **23**, 137–145.

Duffy,M.J. (2005) Predictive markers in breast and other cancers: a review. *Clin. Chem.*, **51**, 494–503.

Dyrskjot,L. *et al.* (2017) Prognostic Impact of a 12-gene Progression Score in Non-muscle-invasive Bladder Cancer: A Prospective Multicentre Validation Study. *Eur Urol*, **72**, 461–469.

Ebell,M.H. (2019) Prolaris Test for Prostate Cancer Risk Assessment. *Am. Fam. Physician*, **100**, 311–312.

Eckhart,L. *et al.* (2013) Cell death by cornification. *Biochim Biophys Acta*, **1833**, 3471–3480.

Elmore,S. (2007) Apoptosis: a review of programmed cell death. *Toxicol Pathol*, **35**, 495–516.

Frenzel,A. *et al.* (2009) Bcl2 family proteins in carcinogenesis and the treatment of cancer. *Apoptosis*, **14**, 584–596.

Fuzio,P. *et al.* (2015) Clusterin transcript variants expression in thyroid tumor: a potential marker of malignancy? *BMC Cancer*, **15**, 349.

Gershenwald,J.E. *et al.* (2017) Melanoma staging: Evidence-based changes in the American Joint Committee on Cancer eighth edition cancer staging manual. *CA Cancer J Clin*, **67**, 472–492.

Ghobrial,I.M. *et al.* (2005) Targeting apoptosis pathways in cancer therapy. *CA. Cancer J. Clin.*, **55**, 178–194.

Global Cancer Observatory.

Gloster,H.M.J. and Neal,K. (2006) Skin cancer in skin of color. *J. Am. Acad. Dermatol.*, **55**, 741–744.

Goossens,N. *et al.* (2015) Cancer biomarker discovery and validation. *Transl. Cancer Res. Vol 4, No 3 (June 2015) Transl. Cancer Res. (Application Genomic Technol. Cancer Res.*

Grill,S. *et al.* (2020) TP53 germline mutations in the context of families with hereditary breast and ovarian cancer: a clinical challenge. *Arch. Gynecol. Obstet.*

Grosso,G. *et al.* (2017) Possible role of diet in cancer: systematic review and multiple meta-analyses of dietary patterns, lifestyle factors, and cancer risk. *Nutr. Rev.*, **75**, 405–419.

Gugnoni,M. *et al.* (2017) Cadherin-6 promotes EMT and cancer metastasis by restraining autophagy. *Oncogene*, **36**, 667–677.

Guha,T. and Malkin,D. (2017) Inherited TP53 Mutations and the Li-Fraumeni Syndrome. *Cold Spring Harb. Perspect. Med.*, **7**.

Guicciardi,M.E. and Gores,G.J. (2009) Life and death by death receptors. *FASEB J.*, **23**, 1625–1637.

Hanahan,D. and Weinberg,R.A. (2000) The hallmarks of cancer. *Cell*, **100**, 57–70.

He,X. *et al.* (2019) POPDC3 is a potential biomarker for prognosis and radioresistance in patients with head and neck squamous cell carcinoma. *Oncol Lett*, **18**, 5468–5480.

Henry,N.L. and Hayes,D.F. (2012) Cancer biomarkers. *Mol. Oncol.*, **6**, 140–146.

Hodgson,A. and Turashvili,G. (2020) Pathology of Hereditary Breast and Ovarian Cancer. *Front. Oncol.*, **10**, 531790.

Ielapi,N. *et al.* (2020) Precision Medicine and Precision Nursing: The Era of Biomarkers and Precision Health. *Int. J. Gen. Med.*, **13**, 1705–1711.

Jin,Z. and El-Deiry,W.S. (2005) Overview of cell death signaling pathways. *Cancer Biol. Ther.*, **4**, 139–163.

Johnson,D.B. *et al.* (2015) Impact of NRAS mutations for patients with advanced melanoma treated with immune therapies. *Cancer Immunol Res*, **3**, 288–295.

Jokinen,E. and Koivunen,J.P. (2015) Bcl-xl and Mcl-1 are the major determinants of the apoptotic response to dual PI3K and MEK blockage. *Int. J. Oncol.*, **47**, 1103–1110.

Kashani-Sabet,M. *et al.* (2017) Prospective Validation of Molecular Prognostic Markers in Cutaneous Melanoma: A Correlative Analysis of E1690. *Clin Cancer Res*, **23**, 6888–6892.

Kazaure,H.S. *et al.* (2018) The impact of age on thyroid cancer staging. *Curr. Opin. Endocrinol. Diabetes. Obes.*, **25**, 330–334.

Kerr,J. *et al.* (2017) Physical activity, sedentary behaviour, diet, and cancer: an update and emerging new evidence. *Lancet. Oncol.*, **18**, e457–e471.

Key,T.J. *et al.* (2020) Diet, nutrition, and cancer risk: what do we know and what is the way forward? *BMJ*, **368**, m511.

Kim,R. (2005) Recent advances in understanding the cell death pathways activated by anticancer

therapy. *Cancer*, **103**, 1551–1560.

Kim,S.-E. *et al.* (2015) Sex- and gender-specific disparities in colorectal cancer risk. *World J. Gastroenterol.*, **21**, 5167–5175.

Koivusalo,M. and Vartiainen,T. (1997) Drinking water chlorination by-products and cancer. *Rev. Environ. Health*, **12**, 81–90.

Koncina,E. *et al.* (2020) Prognostic and Predictive Molecular Biomarkers for Colorectal Cancer: Updates and Challenges. *Cancers (Basel).*, **12**.

Krajewska,M. *et al.* (1996) Elevated expression of Bcl-X and reduced Bak in primary colorectal adenocarcinomas. *Cancer Res*, **56**.

Kretschmer,A. and Tilki,D. (2017) Biomarkers in prostate cancer - Current clinical utility and future perspectives. *Crit. Rev. Oncol. Hematol.*, **120**, 180–193.

Kulasingam,V. and Diamandis,E.P. (2008) Strategies for discovering novel cancer biomarkers through utilization of emerging technologies. *Nat. Clin. Pract. Oncol.*, **5**, 588–599.

Lamb,J. *et al.* (2006) The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science*, **313**, 1929–1935.

Lathwal,A. *et al.* (2020) Identification of prognostic biomarkers for major subtypes of non-small-cell lung cancer using genomic and clinical data. *J. Cancer Res. Clin. Oncol.*

Lee,J.K. and Chan,A.T. (2011) Molecular Prognostic and Predictive Markers in Colorectal Cancer: Current Status. *Curr. Colorectal Cancer Rep.*, **7**, 136–144.

Lever,J. *et al.* (2019) Text-mining clinically relevant cancer biomarkers for curation into the CIViC database. *Genome Med.*, **11**, 78.

Li,B. *et al.* (2020) Expression signature, prognosis value, and immune characteristics of Siglec-15 identified by pan-cancer analysis. *Oncoimmunology*, **9**, 1807291.

Li,J. and Yuan,J. (2008) Caspases in apoptosis and beyond. *Oncogene*, **27**, 6194–6206.

Li,L. *et al.* (1998) [Cellular apoptosis, proliferation and bcl-2 Bax expression in colorectal carcinoma and their association with tumor prognosis]. *Zhonghua Wai Ke Za Zhi*, **36**, 120,614-616.

Li,X. *et al.* (2019) Identification and Validation of Core Genes Involved in the Development of Papillary Thyroid Carcinoma via Bioinformatics Analysis. *Int. J. Genomics*, **2019**, 5894926.

Lian,M. *et al.* (2020) Aging-associated genes TNFRSF12A and CHI3L1 contribute to thyroid cancer: An evidence for the involvement of hypoxia as a driver. *Oncol. Lett.*, **19**, 3634–3642.

Liao,X. *et al.* (2020) WISP1 Predicts Clinical Prognosis and Is Associated With Tumor Purity, Immunocyte Infiltration, and Macrophage M2 Polarization in Pan-Cancer. *Front. Genet.*, **11**, 502.

Liao,Y. *et al.* (2018) Nuclear receptor binding protein 1 correlates with better prognosis and induces caspase-dependent intrinsic apoptosis through the JNK signalling pathway in colorectal cancer. *Cell Death Dis.*, **9**, 436.

Lindner,A. *et al.* (2013) Systems analysis of BCL2 protein family interactions establishes a model to predict responses to chemotherapy. *Cancer Res*, **73**.

Lindner,Andreas U *et al.* (2017) BCL-2 system analysis identifies high-risk colorectal cancer patients. *Gut*, **66**, 2141–2148.

Lindner,Andreas U. *et al.* (2017) BCL-2 system analysis identifies high-risk colorectal cancer patients. *Gut*, **66**, 2141–2148.

Liu,Y.-Q. *et al.* (2019) Gene Expression Profiling Stratifies IDH-Wildtype Glioblastoma With Distinct Prognoses. *Front. Oncol.*, **9**, 1433.

LiVolsi,V.A. (2011) Papillary thyroid carcinoma: an update. *Mod. Pathol.*, **24 Suppl 2**, S1-9.

Loeb,L.A. *et al.* (1984) Smoking and lung cancer: an overview. *Cancer Res.*, **44**, 5940–5958.

Long,G. V *et al.* (2017) Adjuvant Dabrafenib plus Trametinib in Stage III BRAF-Mutated Melanoma. *N Engl J Med*, **377**, 1813–1823.

Luo,G. *et al.* (2020) Roles of CA19-9 in pancreatic cancer: Biomarker, predictor and promoter. *Biochim. Biophys. acta. Rev. cancer*, **1875**, 188409.

Lynch,H.T. *et al.* (2015) Milestones of Lynch syndrome: 1895-2015. *Nat. Rev. Cancer*, **15**, 181–194.

Ma,L. *et al.* (2019) Overexpression of FER1L4 promotes the apoptosis and suppresses epithelial-mesenchymal transition and stemness markers via activating PI3K/AKT signaling pathway in osteosarcoma cells. *Pathol. Res. Pract.*, **215**, 152412.

Maeda,Y. *et al.* (2018) Apigenin induces apoptosis by suppressing Bcl-xl and Mcl-1 simultaneously via signal transducer and activator of transcription 3 signaling in colon cancer. *Int. J. Oncol.*, **52**, 1661–1673.

Mao,Y. and Xing,M. (2016) Recent incidences and differential trends of thyroid cancer in the USA. *Endocr. Relat. Cancer*, **23**, 313–322.

Mathur,P. *et al.* (2020) Cancer Statistics, 2020: Report From National Cancer Registry Programme, India. *JCO Glob. Oncol.*, **6**, 1063–1075.

Mazonakis,M. and Damilakis,J. (2017) Cancer risk after radiotherapy for benign diseases. *Phys. Med.*, **42**, 285–291.

McCann,J.C. and Ames,B.N. (2011) Adaptive dysfunction of selenoproteins from the perspective of the triage theory: why modest selenium deficiency may increase risk of diseases of aging. *FASEB J.*, **25**, 1793–1814.

Meves,A. *et al.* (2015) Tumor Cell Adhesion As a Risk Factor for Sentinel Lymph Node Metastasis in Primary Cutaneous Melanoma. *J Clin Oncol*, **33**, 2509–2515.

Mian,C. *et al.* (1999) Immunocyt: a new tool for detecting transitional cell cancer of the urinary tract. *J. Urol.*, **161**, 1486–1489.

Mittendorf,E.A. *et al.* (2016) The Neo-Bioscore Update for Staging Breast Cancer Treated With Neoadjuvant Chemotherapy: Incorporation of Prognostic Biologic Factors Into Staging After Treatment. *JAMA Oncol.*, **2**, 929–936.

Mor,G. *et al.* (2005) Serum protein markers for early detection of ovarian cancer. *Proc. Natl. Acad. Sci. U. S. A.*, **102**, 7677–7682.

Mughal,F.H. (1992) Chlorination of drinking water and cancer: a review. *J. Environ. Pathol. Toxicol. Oncol.*, **11**, 287–292.

Mumford,S.L. *et al.* (2018) Circulating MicroRNA Biomarkers in Melanoma: Tools and Challenges in Personalised Medicine. *Biomolecules*, **8**.

Musa,A. *et al.* (2018) A review of connectivity map and computational approaches in pharmacogenomics. *Brief. Bioinform.*, **19**, 506–523.

Nakano,T. *et al.* (2020) Overexpression of Antiapoptotic MCL-1 Predicts Worse Overall Survival of Patients With Non-small Cell Lung Cancer. *Anticancer Res.*, **40**, 1007–1014.

Narayanan,D.L. *et al.* (2010) Ultraviolet radiation and skin cancer. *Int. J. Dermatol.*, **49**, 978–986.

Olivier,M. *et al.* (2019) The Need for Multi-Omics Biomarker Signatures in Precision Medicine. *Int. J. Mol. Sci.*, **20**.

Ossio,R. *et al.* (2017) Melanoma: a global perspective. *Nat Rev Cancer*, **17**, 393–394.

Pandya,V. *et al.* (2020) BIK drives an aggressive breast cancer phenotype through sublethal apoptosis and predicts poor prognosis of ER-positive breast cancer. *Cell Death Dis.*, **11**, 448.

Pavlou,M.P. *et al.* (2013) The long journey of cancer biomarkers from the bench to the clinic.

Clin. Chem., **59**, 147–157.

Petrella,A. *et al.* (2006) Annexin-1 downregulation in thyroid cancer correlates to the degree of tumor differentiation. *Cancer Biol. Ther.*, **5**, 643–647.

Qiu,J. *et al.* (2018) Identification of key genes and miRNAs markers of papillary thyroid cancer. *Biol. Res.*, **51**, 45.

Quezada,H. *et al.* (2017) Omics-based biomarkers: current status and potential use in the clinic. *Bol. Med. Hosp. Infant. Mex.*, **74**, 219–226.

Reed,J.C. *et al.* (1996) BCL-2 family proteins: Regulators of cell death involved in the pathogenesis of cancer and resistance to therapy. *J. Cell. Biochem.*, **60**, 23–32.

Reyes,I. *et al.* (2019) Gene expression profiling identifies potential molecular markers of papillary thyroid carcinoma. *Cancer Biomark.*, **24**, 71–83.

Rhea,J.M. and Molinaro,R.J. (2011) Cancer biomarkers: surviving the journey from bench to bedside. *MLO. Med. Lab. Obs.*, **43**, 10–2, 16, 18; quiz 20, 22.

Rich,J.T. *et al.* (2010) A practical guide to understanding Kaplan-Meier curves. *Otolaryngol. Head. Neck Surg.*, **143**, 331–336.

Rinner,B. *et al.* (2004) Activity of novel plant extracts against medullary thyroid carcinoma cells. *Anticancer Res.*, **24**, 495–500.

Samet,J.M. (2013) Tobacco smoking: the leading cause of preventable disease worldwide. *Thorac. Surg. Clin.*, **23**, 103–112.

Sanchez-Vega,F. *et al.* (2018) Oncogenic Signaling Pathways in The Cancer Genome Atlas. *Cell*, **173**, 321-337 e10.

Sasco,A.J. *et al.* (2004) Tobacco smoking and cancer: a brief review of recent epidemiological evidence. *Lung Cancer*, **45 Suppl 2**, S3-9.

Scherr,A.L. *et al.* (2016) Bcl-xL is an oncogenic driver in colorectal cancer. *Cell Death Dis.*, **7**.

Sever,R. and Brugge,J.S. (2015) Signal transduction in cancer. *Cold Spring Harb Perspect Med*, **5**.

Shinwari,Z. *et al.* (2008) Vincristine and lomustine induce apoptosis and p21(WAF1) up-regulation in medulloblastoma and normal human epithelial and fibroblast cells. *J. Neurooncol.*, **87**, 123–132.

Sidransky,D. (2002) Emerging molecular markers of cancer. *Nat. Rev. Cancer*, **2**, 210–219.

Siegel,R.L. *et al.* (2020) Cancer statistics, 2020. *CA. Cancer J. Clin.*, **70**, 7–30.

Sinicrope,F.A. (2018) Lynch Syndrome-Associated Colorectal Cancer. *N. Engl. J. Med.*, **379**, 764–773.

Sivendran,S. *et al.* (2014) Dissection of immune gene networks in primary melanoma tumors critical for antitumor surveillance of patients with stage II-III resectable disease. *J Invest Dermatol*, **134**, 2202–2211.

Smith,A.H. *et al.* (1992) Cancer risks from arsenic in drinking water. *Environ. Health Perspect.*, **97**, 259–267.

Soares,P. *et al.* (2014) Prognostic biomarkers in thyroid cancer. *Virchows Arch.*, **464**, 333–346.

Song,M. and Giovannucci,E.L. (2015) Cancer risk: many factors contribute. *Science*, **347**, 728–729.

Soong,S. *et al.* (2010) Predicting survival outcome of localized melanoma: an electronic prediction tool based on the AJCC Melanoma Database. *Ann. Surg. Oncol.*, **17**, 2006–2014.

Spagnuolo,C. *et al.* (2015) Genistein and cancer: current status, challenges, and future directions. *Adv. Nutr.*, **6**, 408–419.

Stoian,M. *et al.* (2014) Apoptosis in colorectal cancer. *J. Med. Life*, **7**, 160–164.

Szymendera,J.J. *et al.* (1981) Value of five tumor markers (AFP, CEA, hCG, hPL and SP1) in diagnosis and staging of testicular germ cell tumors. *Oncology*, **38**, 222–229.

Tang,Z. *et al.* (2017) GEPIA: a web server for cancer and normal gene expression profiling and interactive analyses. *Nucleic Acids Res.*, **45**, W98–W102.

Todorovic,L. *et al.* (2018) Expression of VHL tumor suppressor mRNA and miR-92a in papillary thyroid carcinoma and their correlation with clinical and pathological parameters. *Med. Oncol.*, **35**, 17.

Toma-Dasu,I. *et al.* (2017) Risk of second cancer following radiotherapy. *Phys. Med.*, **42**, 211–212.

Tomasetti,C. and Vogelstein,B. (2015) Cancer etiology. Variation in cancer risk among tissues can be explained by the number of stem cell divisions. *Science*, **347**, 78–81.

Tong,J. *et al.* (2017) Mcl-1 degradation is required for targeted therapeutics to eradicate colon cancer cells. *Cancer Res.*, **77**, 2512–2521.

Tuli,H.S. *et al.* (2019) Molecular Mechanisms of Action of Genistein in Cancer: Recent Advances. *Front. Pharmacol.*, **10**, 1336.

Visintin,I. *et al.* (2008) Diagnostic markers for early detection of ovarian cancer. *Clin. Cancer Res.*, **14**, 1065–1072.

Vogelstein,B. and Kinzler,K.W. (2004) Cancer genes and the pathways they control. *Nat. Med.*, **10**, 789–799.

Vucicevic,K. *et al.* (2016) Association of Bax Expression and Bcl2/Bax Ratio with Clinical and Molecular Prognostic Markers in Chronic Lymphocytic Leukemia. *J. Med. Biochem.*, **35**, 150–157.

Wang,D. *et al.* (2020) Pan-cancer analysis reveals the role of long non-coding RNA LINC01614 as a highly cancer-dependent oncogene and biomarker. *Oncol. Lett.*, **20**, 1383–1399.

Wang,S.H. *et al.* (2001) Susceptibility of thyroid cancer cells to 7-hydroxystaurosporine-induced apoptosis correlates with Bcl-2 protein level. *Thyroid*, **11**, 725–731.

Wang,S.H. and Baker,J.R. (2006) Apoptosis in thyroid cancer. In, *Thyroid Cancer (Second Edition): A Comprehensive Guide to Clinical Management*. Humana Press, pp. 55–61.

Wang,Y. *et al.* (2018) Identification of a six-gene signature with prognostic value for patients with endometrial carcinoma. *Cancer Med*, **7**, 5632–5642.

Warren,G.W. and Cummings,K.M. (2013) Tobacco and lung cancer: risks, trends, and outcomes in patients with cancer. *Am. Soc. Clin. Oncol. Educ. book. Am. Soc. Clin. Oncol. Annu. Meet.*, 359–364.

Watson,M. *et al.* (2016) Ultraviolet Radiation Exposure and Its Impact on Skin Cancer Risk. *Semin. Oncol. Nurs.*, **32**, 241–254.

Wei,L. *et al.* (2018) TCGA-assembler 2: software pipeline for retrieval and processing of TCGA/CPTAC data. *Bioinformatics*, **34**, 1615–1617.

Weide,B. *et al.* (2012) Serum markers lactate dehydrogenase and S100B predict independently disease outcome in melanoma patients with distant metastasis. *Br J Cancer*, **107**, 422–428.

Weissleder,R. and Ntziachristos,V. (2003) Shedding light onto live molecular targets. *Nat. Med.*, **9**, 123–128.

Wild,C. *et al.* (2015) Cancer risk: role of chance overstated. *Science*, **347**, 728.

Willaert,W. and Ceelen,W. (2015) Extent of surgery in cancer of the colon: is more better? *World J. Gastroenterol.*, **21**, 132–138.

Wu,H.-X. *et al.* (2019) Tumor mutational and indel burden: a systematic pan-cancer evaluation as prognostic biomarkers. *Ann. Transl. Med.*, **7**, 640.

Wu,M. *et al.* (2019) Identification of a Five-Gene Signature and Establishment of a Prognostic Nomogram to Predict Progression-Free Interval of Papillary Thyroid Carcinoma. *Front.*

Endocrinol. (Lausanne)., **10**, 790.

Yang,H.-L. *et al.* (2003) p21 Waf-1 (Cip-1) enhances apoptosis induced by manumycin and paclitaxel in anaplastic thyroid cancer cells. *J. Clin. Endocrinol. Metab.*, **88**, 763–772.

Yang,S.Y. *et al.* (2009) Apoptosis and colorectal cancer: implications for therapy. *Trends Mol. Med.*, **15**, 225–233.

Yi,W. *et al.* (2016) High expression of fibronectin is associated with poor prognosis, cell proliferation and malignancy via the NF-κB/p53-apoptosis signaling pathway in colorectal cancer. *Oncol. Rep.*, **36**, 3145–3153.

Yoshida,R. (2020) Hereditary breast and ovarian cancer (HBOC): review of its molecular characteristics, screening, treatment, and prognosis. *Breast Cancer*.

Yuan,Q. *et al.* (2019) Prognostic and Immunological Role of FUN14 Domain Containing 1 in Pan-Cancer: Friend or Foe? *Front. Oncol.*, **9**, 1502.

Yuan,S. and Akey,C.W. (2013) Apoptosome structure, assembly, and procaspase activation. *Structure*, **21**, 501–515.

Zarkesh,M. *et al.* (2018) The Association of BRAF V600E Mutation With Tissue Inhibitor of Metalloproteinase-3 Expression and Clinicopathological Features in Papillary Thyroid Cancer. *Int. J. Endocrinol. Metab.*, **16**, e56120.

Zeestraten,E.C.M. *et al.* (2013) The prognostic value of the apoptosis pathway in colorectal cancer: a review of the literature on biomarkers identified by immunohistochemistry. *Biomark. Cancer*, **5**, 13–29.

Zeng,S. *et al.* (2019) Prognostic value of TOP2A in bladder urothelial carcinoma and potential molecular mechanisms. *BMC Cancer*, **19**, 604.

Zhao,N. *et al.* (2020) Identification of Pan-Cancer Prognostic Biomarkers Through Integration of Multi-Omics Data. *Front. Bioeng. Biotechnol.*, **8**, 268.