Statistical Learning with Many Variables as Covariates or Outcomes: Association Inference and Prediction of Late effects of Childhood Cancer and Its Treatment

by

Farideh Bagherzadeh-Khiabani

A thesis submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in

Epidemiology

School of Public Health
University of Alberta

ABSTRACT

Advancements in childhood cancer treatment have increased the 5-year survival rates substantially, from 20% in 1950-1954 to over 85% currently. While this success is a remarkable accomplishment in oncology, it concurrently introduces a new concern, namely, the emergence of late adverse effects, commonly referred to as late-effects, of cancer and its treatment in the aging population of long-term survivors. Studies have revealed a diverse spectrum of late-effects experienced by survivors at a much greater extent than the general population of the same ages. The variations in these effects are pronounced, with considerable differences among survivors across individual characteristics, cancer diagnoses, and treatment modalities. Identifying those at higher risks of late-effects and discerning associated factors intensifying this burden are imperative to ensure the lifelong well-being of survivors. This knowledge facilitates targeted interventions tailored to specific high-risk individuals, contributing to the growing recognition of personalized survivorship care.

This dissertation represents a dedicated effort to deepen our understanding of late-effects to enhance cancer survivorship care. Several crucial yet sometimes underappreciated concepts form the core of this dissertation. First and foremost is the advocacy for a holistic view of survivors' experiences, spanning their journey from diagnosis through adulthood, offering potential novel insights into late-effects. Achieving this requires developing and utilizing advanced statistical/machine learning methodologies adept at concurrently handling a spectrum of many covariates, and/or accommodating longitudinal experiences of morbidity that evolve as survivors age.

Second, this dissertation underscores the significance of information directly obtained from survivors, captured in Patient Reported Outcomes (PROs) such as symptoms and health-related quality of life (HRQoL). PROs hold the potential to provide invaluable insights into the future health status and survivorship care needs of survivors, as they contain information known only to the survivors themselves.

Third, this dissertation delves into multi-dimensional health outcomes, such as HRQoL and cumulative count/burden of recurrent, multitype health conditions, aiming to offer a nuanced understanding of the true burden of disease carried by survivors. HRQoL reflects survivors' subjective, multi-dimensional perception of their health and well-being, as affected by disease or treatments. The Mean Cumulative Count (MCC) of recurrent health conditions has been shown to inform true burden of conditions after "cure" over time, surpassing the cumulative incidence of single health conditions. This study seeks to develop a personalized prediction of the cumulative count, moving beyond the conventional group-mean value approach.

This dissertation is structured into three distinct studies, each aimed at gaining fresh insights into the late-effects of childhood cancer and its treatment by addressing one or more of the aforementioned concepts. The first study introduces a novel methodology that facilitates simultaneous consideration of numerous potential covariates during outcome modeling, assessing its performance in covariate selection and coefficient estimation against other alternative techniques through a simulation study. This methodology comprises five key components: generating candidate covariate sets; estimating regression coefficients; scoring candidate models; efficiently searching for candidate models; and enhancing parsimony of the final model.

The second study models HRQoL longitudinally, capturing the dynamic nature of symptoms via advanced statistical/machine learning tools and manually engineered longitudinal patterns. Using the methodology developed in the first study, our findings illuminate key symptom patterns contributing to the longitudinal mental and physical component scores of HRQOL.

The third study introduces a framework for calculating a personalized cumulative disease burden, considering multiple health conditions, potential recurrence, and the competing risk of mortality. This comprehensive approach involves estimating hazard ratios for individual recurrent conditions, estimating hazard ratios for mortality and predicting survival probability, predicting accumulated risk of individual recurrent health conditions, predicting lifelong condition-specific count accounting for the competing risk of mortality, and finally aggregating these counts into an overall burden measure. While we showcase our approach for demonstrating the lifelong burden of multitype chronic health conditions for childhood cancer survivors, this framework can also be utilized to illustrate the longitudinal burden faced by individuals susceptible to any type of recurrent conditions, especially crucial for populations at a heightened risk of mortality.

Collectively, this thesis strives for a comprehensive view towards survivors in measuring late-effects of childhood cancer and its treatment, emphasizing simultaneous evaluation of numerous covariates, consideration of covariate experience through time, inclusion of measures known uniquely to each survivor, and incorporation of multi-dimensional measures of disease burden.

PREFACE

This thesis is an original work by Farideh Bagherzadeh-Khiabani and is part of a broader research project overseen by Prof. Yutaka Yasui. In his role as my PhD advisor, Professor Yasui played a pivotal role in ensuring the ethical integrity of the research and providing guidance on the overall methodological approach employed in this thesis. The project is under the ethnic approval of "Epidemiological analysis of clinical and genetics data" Pro00085976 at University of Alberta.

This work was supported by Grant numbers R01 CA216354, U01 CA195547, U24 CA55727, and R21 CA202210 from the US National Cancer Institute, the American Lebanese Syrian Associated Charities (ALSAC), and the Alberta Machine Intelligence Institute. The data utilized for the analyses in this dissertation was collected by St. Jude Children's Research Hospital through their grant funding provided by the National Cancer Institute, U24 CA055727 "Childhood Cancer Survivor Study" and U01 CA195547 "St. Jude Lifetime Cohort Study".

Chapter 2 of this thesis was submitted as Bagherzadeh-Khiabani F, Huang IC, Martinez-Martinez JM, Izumi S, Dinu I, Yasui Y. "Associational Inference with Many Potential Covariates: Bayesian information Criterion Elastic Net." My responsibilities included developing the methodology, its assessment though a simulation study, as well as result interpretation and manuscript composition, under the supervision of Prof. Yasui. The final manuscript received valuable contributions and critical revisions from all co-authors.

Chapter 3 of this thesis will be submitted as Bagherzadeh-Khiabani F, Krull KR, Armstrong GT, Hudson MM, Robison LL, Yasui Y, Huang IC. "4. Predicting Personalized Burden of Multiple/Recurrent Health Conditions across the Lifespan in Childhood Cancer Survivors." Under

the mentorship of Prof. Yasui, my responsibilities encompassed study design, data analysis, result interpretation, and manuscript composition. The final manuscript benefited from the invaluable contributions and critical revisions provided by all co-authors.

Chapter 4 of this thesis will be submitted as Bagherzadeh-Khiabani F, Daisuke Y, Bhakta N, Yasui Y. “Predicting Personalized Burden of Multiple, Recurrent Health Conditions across the Lifespan: A Case Study on Childhood Cancer Survivors' Chronic Health Burden.” Under the supervision of Prof. Yasui and collaboration with Dr. Daisuke Yoneoka, my responsibilities included designing the study, conducting the analysis, interpreting, and summarizing the results, and composing the manuscript. All co-authors provided valuable contributions and offered critical revisions to the final manuscript. Ms. Qi Liu provided generous assistance in programming and conducting the analysis.

DEDICATIONS

In loving memory of my father, Abolghasem Bagherzadeh-Khiabani.

To Naser, Samyar, and Sepandar.

ACKNOWLEDGEMENTS

I am deeply grateful to many individuals for their substantial contributions to this dissertation. Most notably, I wish to express my profound appreciation and respect for my mentor, Prof. Yutaka Yasui. The wealth of knowledge I have gained from him over the years is immeasurable. Prof. Yasui has been an extraordinary mentor, offering unwavering support, not only throughout my academic journey but also beyond. I am thankful for the invaluable constructive feedback and guidance he provided during the development of this dissertation. Collaborating closely with him has been a true privilege.

I also extend my sincere gratitude to the esteemed members of my PhD Supervisory Committee, Prof. Shelby Yamamoto, Prof. Sentil Senthilselvan, and Prof. Irina Dinu, for their invaluable feedback. Their collaborative efforts have been integral to this work, and I am deeply appreciative of their contributions to my professional development. I would also like to express my thanks to the dedicated professors at the School of Public Health, as well as the faculty members and researchers at St. Jude Children's Research Hospital.

My research received support from multiple institutions and funding agencies. I am especially grateful to St. Jude Children’s Research Hospital for granting me access to exceptional research resources. Additionally, I would like to express my gratitude to the Alberta Machine Intelligence Institute for their financial support throughout my entire PhD program.

Lastly, it is impossible for me to adequately convey the profound appreciation I hold for my family. Undertaking this educational journey would have been unthinkable without their support. To my

husband, Naser Abdollahi – I want to extend my heartfelt gratitude for your enduring patience and unconditional support.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1. Background and significance of late-effect research in childhood cancer survivors

Despite significant investment in childhood cancer research to enhance our understanding of cancer and improving its treatments, there remains a lot to be learned about health of childhood cancer survivors following their active treatment. The following reasons highlight the need for increased research in childhood cancer survivorship research.

1.1.1. An expanding population

The number of childhood cancer survivors has been steadily increasing worldwide. Annually, more than 400,000 new children are estimated to be diagnosed with cancer in the world, and projections indicate a staggering 13.7 million new cases between 2020 and 2050. A noticeable disparity exists in the diagnosis, treatment, and care provided to children with cancer across various countries worldwide [1]. In the US, where statistics are readily available and mirroring the situation in Canada, there were an estimated 500,000 childhood cancer survivors in 2020, up from around 330,000 in 2005 [2, 3]. It is projected that in 2023, about 15,190 children and adolescents aged 0-19 were estimated to have been diagnosed with cancer in the US, which translates to roughly 42 new cases every day, of whom over 85% will become 5-year survivors. Approximately, in 2023, 1 in every 260 children and adolescents in the US were estimated to have a history of cancer [4].

With survivors of childhood cancer continuing to grow into adulthood, childhood cancer survivorship studies are increasingly becoming a matter of public health concern, and focusing on

the health and well-being of this large and ever-increasing population of cancer survivors is imperative [5].

1.1.2. A high-risk population

Survivors of childhood cancer are at increased risk of morbidity, premature mortality, and reduced quality of life associated with their cancer treatments. In a study of 5,522 5-year survivors treated for childhood cancer at St. Jude Children's Research Hospital and 272 controls, Bhakta et al. observed that, at age 50, 99.9% of survivors (vs. 96.0% of controls) experienced at least one chronic condition and 96.0% of survivors (vs. 84.9% of controls) experienced at least one severe/disabling, life-threatening, or fatal conditions. By age 50, a survivor had experienced, on average, 17.1 chronic health conditions of any grade, of which 4.7 were severe/disabling, life-threatening, or fatal conditions (vs. 9.2 and 2.3 for controls, respectively). These values indicate the burden of morbidity to be twice for survivors compared to controls [6]. A study of 1,667 10-year survivors of childhood cancer who participated in St. Jude Lifetime Cohort Study (SJLIFE) also indicated more than 75% of survivors experienced multiple symptoms, and showed a higher symptom burden compared to the general population [7]. Using data from the Childhood Cancer Survivor Study (CCSS), studies showed that survivors live approximately 4 to 18 fewer years than the general population, a life expectancy reduction of up to 28% [8]. Also, a study of childhood cancer survivors from Surveillance, Epidemiology, and End Results (SEER) registries estimated 126,952 excess deaths (compared to the age, sex, race, and calendar-year matched US population) in the estimated population of 445,647 US individuals under 20 years of age, diagnosed with cancer between 1975 and 2016. The nearly flat trend of the estimated annual excess death of survivors over diagnosis year, observed in this study, implies the persisting high burden of disease in childhood cancer survivors in spite of the achieved success in their life-extension [9].

1.1.3. Decades with morbidity

Although constituting less than 5% of the total population of cancer survivors, survivors of childhood cancer are of particular importance for the following reasons. First, these survivors undergo treatments during crucial periods of physical, developmental, and psychological growth which could affect growing and developing tissues and impede the attainment of developmental milestones. Second, following diagnosis at a young age, survivors confront a long survivorship, often spanning six decades or more, living with morbidity. This long survivorship is accompanied with longer periods of care needs from families and care givers, and thus complicating their lives. The longevity of survivorship also poses a high economic burden on the society due to the medical costs, and the lower unemployment and underemployment levels of childhood cancer survivors over a lifetime, compared to similar individuals without cancer, even many years after diagnosis. Finally, when working with young individuals with decades of life ahead of them, long-term consequences of therapy could be as important to consider as acute and short-term consequences of cancer, leading to potentially prioritizing treatments that offer better overall health outcomes in the long run, even if they come with a slightly increased risk of acute/early effects [5, 10-12].

1.1.4. A dynamic population with a diverse experience

Childhood cancer survivors constitute a distinctive group of individuals due to their unique and diverse experience. Firstly, cancer is a diverse disease, with experiences that vary widely depending on factors such as tumor's classification (origin of the tumor cells), grade (degree of malignancy), and stage (extent of tumor spread). Secondly, treatment options for childhood cancer are multifaceted and may consist of new combinations of chemotherapy agents, surgery, radiation therapy, or an entirely new approach like immunotherapy, and are continuously evolving (as new information about late-effects of treatment emerges, therapies for childhood cancer are being

modified to minimize the likelihood of late-effects), which could lead to the emergence of new late-effects. Thus, the population of cancer survivors could be thought of as a moving target. Thirdly, the experience of late-effects could be highly variable in nature, with some being identified immediately and then resolve without consequence, and others persisting for long periods, or only becoming clinically apparent years after treatment, necessitating long and complete follow-up of cancer survivors. Finally, due to the relatively new access to a long-term followed population of childhood cancer survivors, there remains a lot to learn about the overall burden of morbidity in childhood cancer survivors as they age and deal with other comorbid conditions that arise [13, 14].

Considering the aforementioned points, it is imperative to undertake public health research with the ultimate objective of improving survival rates, the overall well-being, and quality of life of childhood cancer survivors.

1.2. Impact of late-effect research in childhood cancer survivors

Cancer survivorship research holds promise for improving the lives of current and future generations of childhood cancer survivors, providing information that could be used to enhance evidence-based clinical practice guidelines and inform shared decision-making by survivors and clinicians, throughout the survivorship continuum from initial treatments to survivorship care into adulthood.

Well-designed cancer survivorship research has the potential to inform and improve future clinical practice guidelines by identifying which survivors are likely to develop late-effects and/or the risk factors underlying these late-effects. Specifically, identifying high-risk survivors, i.e., those with a high burden of morbidity and reduced quality of life, could help target the right population,

ensuring that appropriate intervention and screening strategies are employed to benefit those who need it and benefit from it the most. Also, elucidating mechanisms that put survivors at high risk could help guide the development of interventions [5, 12-17].

Integrating user-friendly decision-making tools such as risk calculators, derived from cancer survivorship research, into survivorship care plans or personal portals can assist survivors in scheduling timely medical appointments, and can facilitate helpful communication between primary care physician and survivor about specific late-effects and needs of survivors that the physician might have not been aware of [18].

Cancer survivorship research could help with critical decisions both during the active treatment period of cancer (i.e., primary prevention of late-effect by replacing a treatment by another with lower risk of late-effect when both appearing the same regarding survival) and during the post-treatment cancer survivorship care (i.e., secondary prevention of late-effects by providing counselling and other interventions to those with the highest risk of poor outcomes).

1.3. Limitations of current late-effect studies in childhood cancer survivors

1.3.1. Limited application of advanced statistical/machine learning tools for data-driven knowledge discovery in the context of late-effects

With the widespread recognition of the importance of risk-based care for cancer survivors to mitigate the late-effects, there has been a growing demand for research to address this need [5, 16, 17]. As conventional in health studies, studies of late-effects in childhood cancer survivors mostly start with an a priori hypothesis about covariate-outcome relationships and utilize classic statistical tools such as regression to test the hypothesis. However, it is increasingly challenging to formulate a clear hypothesis on covariate-outcome relationship for the wide spectrum of late-effects due to the heterogeneity of cancer survivor population and their exposures, continuously changing

treatments, and the increasing number of newly collected covariates (such as patient-generated symptoms and vital signs via wearable or mobile devices). Further, conceptually, it is becoming widely acknowledged that health status of a survivor in the long run are influenced by a multitude of factors (and their complex interplay) that are not necessarily directly linked to cancer and its treatment (see Figure 1.1): for example, if a person is less capable of adjusting to their unexpected complications, this needs to be remembered in assessing their long-term health status [5, 16, 19]. As such, state-of-the-art statistical/machine learning tools that are free from a priori assumptions and start from a wide spectrum of possible risk factors are highly relevant in cancer survivorship research, and offer the promise of unraveling novel insights and timely breakthroughs, such as discovery of new risk factors relevant to prognosis and pathology, in the relatively new long-term follow-up cancer survivor data. The potential for revolutionizing cancer survivorship care through the combination of machine learning, high-dimensional data, and computational power has yet to be fully realized.

1.3.2. Gap between methodological advances for prediction & needs for epidemiological association inference in the presence of many candidate covariates

Although machine/statistical leaning techniques have made salient progress in developing methodologies for outcome prediction in the presence of many predictors/covariates, there has been comparatively less progress in developing methodologies for association inference. Association inference involves finding an accurate and clear explanation of outcomes obtainable from data [20-23]. As such, in practice, association inference in biomedical research often relies on subject-matter knowledge/biological meaningfulness of covariates, data-driven univariate analysis, and/or machine/statistical learning methods designed intrinsically for outcome prediction rather than inference [23, 24].

The common practice of using subject-matter knowledge or biological meaningfulness to choose covariates could be challenging when the number of potential covariates is large and/or there is limited prior knowledge about them. This approach also limits the possibility of discovering novel covariate sets that might describe the data mechanism better than the pre-specified ones [25, 26]. Similarly, the widely used practice of data-driven univariate tests to screen variables could disregard covariates that are truly associated with the outcome in conjunction with other covariates but not univariately [27]. The trending practice of employing prediction-focused methods from machine/statistical learning for inference is also concerning because although good prediction models should asymptotically represent the true associations, with a finite sample, they may not, which could lead to incorrect inferential conclusions. Therefore, it is critical to design state-of-the-art association-inference methods that prioritize the needs of health researchers such as the familiarity of methodological concepts.

1.3.3. Ignoring survivors' voice

Due to different priorities and values, and coping mechanisms of each survivor, it is important to consider their personal perspective on their well-being, including how well they perceive themselves doing as survivors in their adult life given their unique factors such as their ability to cope. Survivors themselves are an invaluable source of information when it comes to reporting the breadth of issues arising during their cancer journey, and their perspectives should be given serious consideration in survivorship research. Survivors are likely to prioritize quality over quantity when estimating their disease burden, providing effective information on improving their quality of care.

PROs provide an ideal method for capturing the subjective nature of the side effects and late-effects experienced by cancer survivors. By emphasizing what survivors feel and report, PROs can help in the development of patient-centered care [28, 29]. Survivors' perceived health status,

measured by PROs, have been shown to predict the onset of chronic health conditions and subsequent survival. A growing body of evidence suggests that regular collection of PROs from cancer survivors resulted in better QOL after six months, reduced emergency room visits after 36 months, and improved survival rates between one and eight years compared to standard care without PRO collection [30-32]. Survivors' symptom burden have been shown to be associated with QoL [7]. PROs could trigger timely clinical evaluation and timely decisions, while also potentially reducing the burden of unnecessary clinic visits for survivors.

Despite this widespread recognition of the importance of PROs [33], they often remain supplementary in clinical trials, survivorship or primary care visits, and statistical analysis. Clinical trials typically focus on PRO measures only as a secondary or exploratory endpoint [14]. Further, during in-person visits, clinicians often focus on clinical measures such as risk factors for chronic conditions or recurrence, neglecting to ask about the survivors' own input on their conditions [34]. Statistical analysis efforts for risk assessment for late-effects also tend to focus on demographic and treatment variables [35]. To truly deliver patient-centered healthcare, it is essential to enhance the integration of PROs into cancer survivorship research.

1.3.4. Overreliance on survivor reports of clinical outcomes without medical validation

Studies on cancer survivors often rely on self-reported data of chronic health conditions (CHC) without clinical verification. However, this approach has limitations and may result in under- or over-reporting of some CHCs [36]. A study using patient-reported CHCs found that 88% of childhood cancer survivors aged 40-49 years had at least one CHC (grades 1-4), including the 48% estimated to have a serious/disabling or life-threatening chronic condition (grades 3-4) [37]. However, a second study using clinically assessed CHCs suggested that these values were

underestimated and estimated 95.5% (vs. 88% above) and 80.5% (vs. 48% above) of survivors having grades 1-4 and grades 3-4 CHCs, respectively, at age 45 years [38]. Thus, integrating clinically assessed outcomes into cancer survivorship research could offer an opportunity to refine our understanding of the burden of cancer and its treatments.

1.3.5. Considering symptoms as fixed characteristics of survivor and reliance on a priori hypothesized clusters of symptoms

Following the revealed importance of symptoms as associative and predictive factors of health outcomes, there are increasing publications on integrating them into cancer survivorship as well. While this relatively new trend has successfully confirmed symptoms' importance in cancer survivorship research, there are some limitations. Firstly, relying on symptoms at a snapshot of time assumes symptoms are fixed characteristics of survivors. However, symptoms could fluctuate over time due to various factors such as progression of co-morbid conditions, aging, the accumulation of treatment effects. The dynamic nature of individual symptoms could better be addressed by a longitudinal approach to monitoring, collection and analysis of symptoms. Additionally, relying on few a priori hypothesized symptom patterns may limit the exploration of novel and complex ways that symptoms could inform about health status. This could better be addressed by considering various patterns of symptom occurrence through time and/or allowing for data-driven selection of combination/co-occurrences of symptoms of different nature. Therefore, adopting a longitudinal approach to studying symptoms, considering patterns of their existence through time, and allowing for novel combinations of factors could provide valuable insights into survivors' health status.

1.3.6. Focusing on single late-effects and first occurrences

Survivors of childhood are at a lifelong risk of several chronic health conditions. Previous cancer survivorship research is mostly focused on one CHC at a time and using cumulative incidence

measures that only account for the first occurrence of the late-effect condition. The study by Bhakta et al., estimating cumulative burden of morbidity, shows that while similar proportions of survivors and controls (99.9% and 96.0%, respectively) experienced at least one CHC by age 50, the burden of morbidity, measured by mean cumulative count of the number of CHCs, was almost twice for survivors compared to controls (17.1 and 9.2, respectively) [6]. These values suggest the importance of comprehensive measures of disease burden that account for multiple late-effects, recurrences of late-effects and death in order to truly show the true price of cancer and its treatment in cancer survivors (a sub-group at high-risk for many late-effects and death). While the index is increasingly used to assess the excess magnitude of late-effects in cancer survivors relative to controls and to assess the corresponding associative factors, further research is needed to predict the individual-level cumulative count of CHCs for cancer survivors.

1.4. Research questions and specific aims

1.4.1. Overall objectives

The overarching objective of this dissertation is to contribute to the understanding of the complex interplay of risk factors affecting late-effect outcomes in childhood cancer survivors. This endeavor is pursued by leveraging advanced statistical/machine leaning tools, reducing reliance on existing subject-matter literature, and making use of novel data sources and innovative outcome measures. Innovations in the analysis of long-term childhood cancer survivor data could provide more accurate and individualized estimates of the magnitude of late-effect risks and reveal new insight into the combination of risk factors and possible pathways that contribute to the late-effects.

1.4.2. Specific aims

The specific aims for the proposed study are as follows:

- **Specific Aim 1 (Method Development):** Develop a statistical tool for epidemiological association inference and evaluate its performance versus existing methods in the presence of numerous potential risk factors.
- **Specific Aim 2 (Method Application):** Examine associations of hundreds of longitudinal patterns of 37 patient-reported symptoms over 3 time points with future HRQoL in survivors of childhood cancer, infer a subset of truly-associated patterns, and estimate their HRQoL associations.
- **Specific Aim 3 (Method Development):** Develop a statistical model for the cumulative count of multitype/recurrent health conditions over time with competing risk that depends on the history of health conditions.

1.5. General methods

1.5.1. Data Resources

Childhood Cancer Survivor Study (CCSS) and St. Jude Lifetime Cohort Study (SJLIFE) provide invaluable resources for cancer survivorship research. Their comprehensive and detailed data, collected over a prolonged period longitudinally with exceptional precision, provide a unique opportunity to characterize and understand the full extent of late-effects associated with childhood cancer and its treatment. CCSS and SJLIFE are both retrospectively-constructed cohorts with prospective follow-up of childhood cancer survivors who survived at least 5 years post cancer diagnosis, designed for assessing late-effects of childhood cancer and its treatment. CCSS, initiated in 1994, follows survivors diagnosed between 1970 and 1999 who had been treated for childhood cancer at 31 institutions in North America, and their siblings who did not have childhood serious illnesses as the comparison group. Including an extensively characterized cohort, CCSS facilitates addressing a wide range of questions. Information on late-effects, in CCSS, have been collected

through periodic comprehensive questionnaires. SJLIFE, initiated in 2007, follows survivors diagnosed after 1962 who had been treated at the St. Jude Children's Research Hospital (SJCRH), and community controls aged at least 18 years old. With the aim to facilitate longitudinal comprehensive clinical evaluation of health outcomes of a lifetime cohort of adult survivors from childhood cancers, all SJLIFE participants have been assessed longitudinally through multiple-day visits to SJCRH where they undergo medical, physical, psychosocial, and neurocognitive assessments. Details of the studies have been published previously [39-41].

1.5.2. Analytic Methods

Machine learning modelling techniques

Throughout this research endeavor, our intention is to leverage machine learning tools for the extraction of knowledge from extensive datasets. Machine learning modelling techniques employ mathematical procedures to discern patterns within datasets, making minimal assumptions about the relationship between covariates and outcomes. In the context of healthcare, the application of machine learning holds the potential to transform medical practices by advancing our understanding of health outcomes (such as identifying risk factors for late-effects of cancer and its treatments) and predicting future state of health outcome for individuals (such as predicting survivors' risk of late-effects). The transformative potential of machine learning in medicine lies in its capacity to leverage real-time information from vast patient datasets to enable personalized diagnoses, management decisions, and therapies for an individual patient. This undertaking could prove overwhelming and impossible for individual physicians, often leading them to choose the most familiar course of action rather than the potentially more effective decision that could be made based on the latest available evidence [42, 43].

Distinguishing machine learning from statistical learning involves recognizing that there is no clear distinction between the two. Machine learning models and traditional statistical models exist along a continuum based on the extent to which their structures or parameters are predetermined by humans: as assumptions decrease, we move away from statistical tools and toward machine learning tools. It is important to note that machine learning is not a miraculous tool capable of turning any data into useful information; rather, it just serves as an extension of conventional statistical methodologies. The vast and intricate landscape of medicine, coupled with the abundance of thousands of patient measures, can make formulating assumptions challenging, which contributes to the growing popularity of machine learning tools in this field [44].

Two modelling cultures: association inference modelling & risk prediction modelling

During this research undertaking, our goal is to formulate models for explaining and predicting late-effect outcomes in survivors of childhood cancer. It is essential to underscore that both association modeling and prediction modeling, prevalent in statistical/machine learning contexts, hold scientific significance in biomedical research, each addressing distinct research questions and concentrating on specific aspects. Prediction models play a pivotal role in identifying high-risk subgroups for targeted interventions and suggesting personalized treatments to patients and physicians. On the other hand, association models serve as valuable tools for delivering explanations of outcomes derived from data, and thereby contributing to a deeper understanding of health outcomes.

Aligned with their distinct objectives and applications, prediction models prioritize predictive performance while association models prioritize explanatory power. Furthermore, these models yield varying types of measures in accordance with their intended use. Prediction models, operating at the individual level, provide absolute risk scores for individual patients. In contrast,

association models, functioning at the group level, provide average relative risk measures, such as relative risk, facilitating comparison of risk factor subgroups with regards to the outcome and enabling covariate-outcome association investigation. A famous example of prediction model in healthcare is the Framingham risk score, and a famous example for the association modelling is genome-wide association studies [20-23].

Feature engineering as a preprocessing step

Feature engineering involves conversion of original covariates into an alternative representation deemed more informative for the model's specific objective. For instance, transforming combinations of covariates into new variables can sometimes prove more effective than using the original covariates. The optimal feature engineering strategy often relies on the problem-specific understanding [45]. In this dissertation, within the section focused on utilizing symptom covariates for modelling HRQoL, we employed feature engineering techniques to transform the symptom data obtained from three symptoms surveys conducted over a span of 20 years into longitudinal symptom patterns. In our case, the decision to transform the original symptom data to longitudinal symptom patterns was made in acknowledgment of the dynamic nature of symptoms throughout the survivorship journey, aiming to capture patterns of symptom presence over time as more stable characteristics of survivors.

Scoring and performance measures utilized

In alignment with the objectives for each section of this dissertation, we employed a range of scoring and performance measures to compare and evaluate the developed methods/models. The Bayesian Information Criterion (BIC), the negative log-likelihood of the model penalized for model's complexity, was used as a scoring measure to rank the many candidate models due to reasons such as ease of computation and its ability to compare non-nested models. To assess the

performance regarding outcome prediction, we employed the Area Under the Curve (AUC) for binary-outcome prediction and Mean Squared Error (MSE) for continuous-outcome prediction both on out-of-training sample data. The AUC, a widely used metric in health studies, reflects probability that the model assigns a higher probability to a randomly chosen patient with the outcome than to a randomly chosen patient without the outcome. The Mean Squared Error measures the average squared difference between predicted and actual outcomes. To assess analytic methods regarding the selected covariates and estimated coefficients, we employed two agreement measures: the Jaccard Similarity Coefficient (JSC) for covariate set assessment and the Intra-Class Correlation Coefficient (ICC) for coefficient estimate assessment. The JSC is defined as the ratio of the number of covariates shared between two compared sets (intersection) to the number of covariates present in either set (union). The ICC could be defined as the ratio of the variance attributed to differences between the coefficient estimates of distinct covariates in the two compared vector to the total variance attributable to both the between-covariate and within-covariate coefficient estimates.

Resampling techniques employed

This dissertation employs two resampling techniques, namely cross-validation and bootstrapping, to facilitate the reporting of the aforementioned measures. In cross-validation, the dataset is partitioned into k non-overlapping folds, with the model iteratively trained on k-1 folds and tested on the remaining one. The cycle repeats k times, ensuring each fold serves as a testing set exactly once. The model's overall predictive performance is subsequently determined by averaging the results across all test folds. Cross-validation is particularly valuable for reporting unbiased predictive performance of a model. In bootstrapping, multiple random samples are repeatedly drawn with replacement from the original dataset, followed by the application of the statistical

analysis to each bootstrap sample. Bootstrapping is particularly valuable for investigating the stability of a modelling technique.

Penalized regression methodologies for covariate/feature selection

Due to the considerable number of covariates across all sections of this dissertation, we employ covariate selection techniques to avoid overfitting. Here, we elaborate on covariate selection in general and provide a brief explanation of penalized regression technique as the primary tool used in all three sections of this dissertation. We then describe two popular penalized techniques utilized in this dissertation, namely Least Absolute Shrinkage and Selection Operator (LASSO) and Elastic Net (EN).

- **Covariate selection:** Covariate selection involves identifying and retaining the most relevant covariates from a pool of potential covariates for the purpose of explaining and predicting an outcome [46, 47]. Effective covariate selection in health studies ensures that models not only lead to better performance but are also more interpretable and robust.
- **Penalized regression:** Penalized regression is a regression method renowned for its resistance to overfitting. This is achieved through an optimization function that penalizes the goodness of fit of models for their complexity. The complexity, measured in terms of the sizes of regression coefficients, is integrated into a single optimization function that considers both the coefficient parameters and hyper-parameters controlling the extent and shape of penalization [47, 48]. The candidate covariates in the initial pool simultaneously are all assessed simultaneously in this optimization function. This joint evaluation allows for the selection of covariates based on their collective impact, a crucial aspect in biomedical research studies.

- **LASSO:** The LASSO optimization function, in case of linear regression, takes the following form:

$$\min_{\beta_0,\beta}\left\{\frac{1}{n}\sum_{i=1}^{n} l(y_i, \beta_0 + \sum_{j=1}^{p}\beta_j x_{ij}) + \lambda \sum_{j=1}^{p}|\beta_j|\right\} \quad (1)$$

where n is the number of subjects, l is the negative log-likelihood, p is the number of covariates, y_i is the outcome for the i[th] subject, β_0 is the intercept, β_j is the regression coefficient of the j[th] covariate, x_{ij} is the value of the j[th] covariate for the i[th] subject, and λ (≥0) is the hyper-parameter determining the extent of penalization balancing the goodness of fit versus model complexity. Group LASSO with overlap, an extension of LASSO, enables the selection of predefined overlapping covariate groups. This approach allows for the simultaneous inclusion/exclusion of entire covariates groups, offering the possibility to select an interaction term into the model only in the presence of its main effects [49-52].

- **Elastic Net:** The Elastic Net optimization function, in case of linear regression, is as follows:

$$\min_{\beta_0,\beta}\left\{\frac{1}{n}\sum_{i=1}^{n} l(y_i, \beta_0 + \sum_{j=1}^{p}\beta_j x_{ij}) + \lambda\left[\alpha \sum_{j=1}^{p}|\beta_j| + (1-\alpha)\sum_{j=1}^{p}\beta_j^2/2\right]\right\} \quad (2)$$

where the additional α (0≤α≤1) hyper-parameter is determining the shape of the penalization term through weights assigned to the sum of squared and the sum of absolute values of the regression coefficients. Compared to LASSO, Elastic Net provides more stable variable selection in the presence of multicollinearity but comes with increased computational demands [53].

1.5.3. Innovative approaches

Bayesian Information Criterion Elastic Net (BIEN) framework

The pursuit for uncovering meaningful associations in biomedical research in the presence of numerous potential covariates has motivated us to develop and utilize a methodology, referred to as the Bayesian Information Criterion Elastic Net (BIEN), offering biomedical researchers an efficient and familiar means of analysis. We briefly introduce BIEN through its five components:

- **Component 1:** To start the process, BIEN employs Elastic Net (EN) to generate potential covariate sets for subsequent selection. The creation of candidate covariate sets involves the exploration of numerous combinations of the two hyperparameters within the optimization function of the Elastic Net.
- **Component 2:** BIEN leverages Maximum Likelihood Estimation (MLE) for the estimation of regression coefficients in candidate models. We chose MLE for its consistency under a set of regularity conditions [54, 55], making it particularly appealing for association inference. In contrast, Elastic Net, while proficient in prediction, may yield inconsistent estimates for parameter estimation and inference [56, 57].
- **Component 3:** The scoring of fitted candidate models in BIEN is accomplished through the BIC. Similar to EN, BIC is a negative log-likelihood value penalized for model complexity. This consistency in penalization allows for a seamless integration of BIC into the methodology. BIC's appeal to be integrated into the methodology lies in its consistency [58], approximation to Bayes Factor [59], compatibility with Bayesian logic, ease of computation without the need for priors, and the ability to compare non-nested models [59].
- **Component 4:** BIEN employs a truncated grid search to find the optimal values for the hyper-parameters (α and λ) that produce models with optimal BIC scores. The search encompasses a

range of plausible values for α and λ, both starting from zero but the former going up to 1 and the latter having an upper limit defined based on the data at hand in practice [60]. The search is strategically truncated to improve computational efficiency, avoiding unnecessary complexity, and preventing over-optimistic BIC values. The introduction of a "patience" hyper-parameter further refines the search process, mitigating the risk of identifying local optima as global optima.

- **Component 5:** Once a model has been selected based on the aforementioned four components, its parsimony is increased using a backward elimination process. This step starts with the selected MLE model and involves removing the most uninformative covariates one by one until BIC score stops improving.

Personalized Cumulative Prediction (PCC) framework

In the pursuit for advancing personalized medicine, we propose and utilize an innovative framework aimed at predicting a personalized health-related burden, referred to as the PCC. This framework endeavors to quantify the age-specific cumulative count of recurrent, multitype health-related conditions throughout an individual's lifespan, taking into consideration their unique characteristics. We briefly introduce this framework through its five steps:

- **Step 1:** In the initial step, our framework involves estimating hazard ratios and the baseline hazard for each specific recurrent health condition of interest individually, given its relevant predictors. This step could easily be achieved through conventional statistical tools.
- **Step 2:** Moving to the second step, recognizing the crucial role of mortality as a competing-event risk in predicting the marginal count of health conditions - the focal task for step 4 – the objective of Step 2 is to estimate hazard ratios for mortality to facilitate the prediction of survival probabilities within the specified lifespan of interest. The mortality model accounts

for relevant demographic, treatment covariates, and the cumulative number of recurrent health conditions, treating the latter as a time-varying covariate representing the longitudinal experience of morbidity.

- **Step 3:** For each recurrent health condition, the accumulated risk over the lifespan of interest is derived by aggregating the corresponding instantaneous risk associated with that condition.
- **Step 4**: Building upon the values acquired in the initial three steps, we proceed to predict the cumulative count of each recurring health condition throughout the lifespan of interest, denoted as condition-specific PCC
- **Step 5:** Finally, the condition-specific PCCs are aggregated to yield an overall PCC that takes into account multitype conditions.

1.6. Ethics statement

Institutional Review Board approval for the proposed analysis has been received by St. Jude Children's Research Hospital.

1.7. Dissertation structure

The subsequent chapters in this dissertation will provide an in-depth exploration of the three papers this dissertation is based on. Chapter 2 will delve into a cutting-edge methodology for inference in high-dimensional data and findings of Paper 1. Chapter 3, utilizing the methodology proposed in chapter 1, will attempt to model several components of HRQoL, including mental and physical component scores, in childhood cancer survivors with a focus on longitudinal symptom patterns as predictors, and discuss findings of paper 2. Chapter 4 will provide a detailed account of the personalized burden metric framework, presented in Paper 3, to better quantify the health burden faced by childhood cancer survivors through their lifespan. The concluding chapter, Chapter 5, will synthesize the findings of the three papers, tying them together into a cohesive narrative. It

will also discuss the broader implications of our research, including the public health significance and the potential to influence the development of personalized healthcare strategies for childhood cancer survivors. Through this dissertation, we aim to contribute to the ongoing discourse on late-effects in childhood cancer survivors, ultimately offering a more nuanced and personalized approach to healthcare for this growing population.

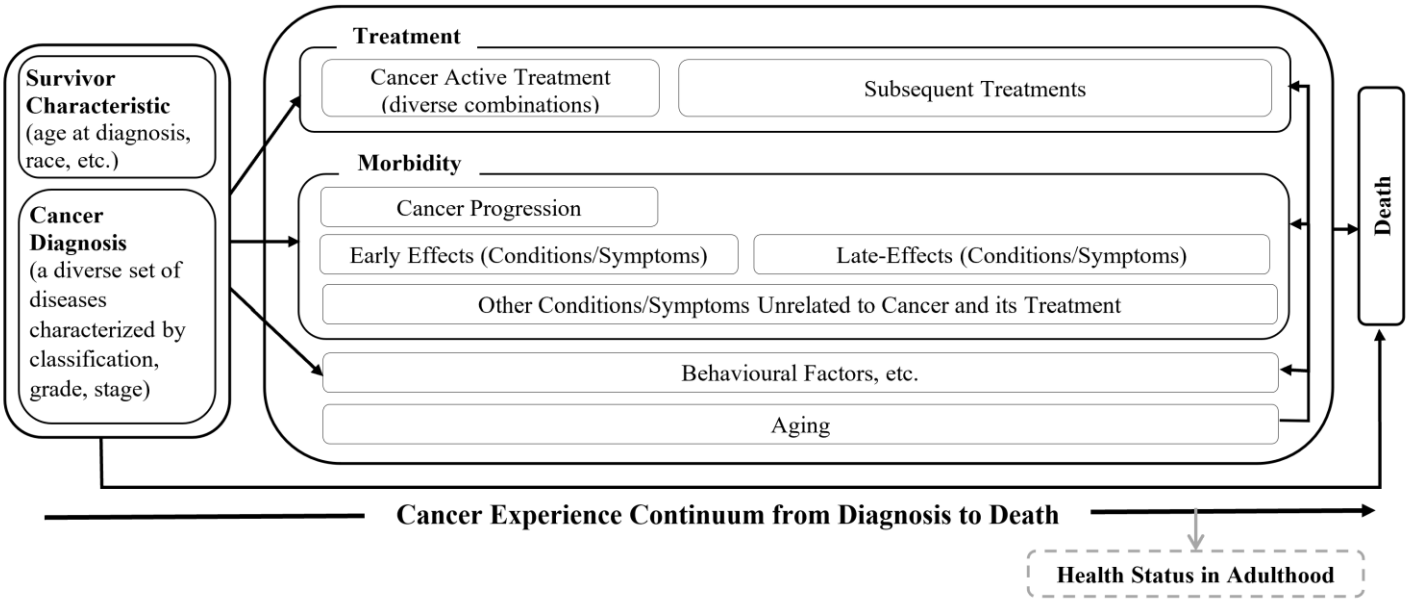


Figure 1.1: Cancer experience continuum.

CHAPTER 2

Associational Inference with Many Potential Covariates: Bayesian information Criterion Elastic Net

2.1. Introduction

With continued advancements in computer and measurement technologies, the number of covariates (explanatory variables in regression) collected in research and surveillance continues to grow. In biomedical research, such examples include the routine collection of patient-generated symptoms and vital signs via wearable or mobile devices, as well as the measurement of germline DNA sequences from blood and buccal cell samples. While a large number of covariates provides researchers with new opportunities for data-driven hypothesis generation and testing, making inference on covariates from the ever-expanding pool of candidates is increasingly challenging, because it involves both selecting relevant covariates and estimating/hypothesis testing on their parameters (regression coefficients).

Advanced methods from machine/statistical learning have been successful in developing tools for predicting outcomes in the presence of many predictors/covariates. These tools are useful in biomedical research for identifying high-risk subgroups and risk-stratifying patients for risk-specific intervention approaches. Alternatively, scientific interest may be finding an accurate and clear explanation of outcomes obtainable from data, which is the goal of association inference. That is, in such scenarios, researchers are interested in *associations of the outcome with covariates*, their presence/absence and the degrees of the associations, not predicting the outcome with covariates. While prediction and association inference are both of scientific interest, they address distinct research questions and should not be practiced indiscriminately [20-23].

Association inference in the presence of many potential covariates is an increasingly common scenario in biomedical research. There exists a critical gap, however, between the available data-analytic methods and applications [26]. Association inference in biomedical research often relies on subject-matter knowledge/biological meaningfulness of covariates, data-driven univariate analysis, and/or machine/statistical learning methods designed intrinsically for outcome prediction rather than inference [23, 24]. The common practice of using subject-matter knowledge or biological meaningfulness to choose covariates could be challenging when dealing with a large number of potential covariates and/or limited prior knowledge on them. This approach also limits the possibility of discovering novel covariate sets that might describe the data mechanism better than the pre-specified ones [25, 26] such as biological pathways linking many molecular covariates. Also, the widely used practice of data-driven univariate tests to screen variables could disregard covariates that are truly associated with the outcome in conjunction with other covariates but not univariately [27]. Furthermore, the trending practice of employing prediction-focused methods from machine/statistical learning for inference raises concerns: asymptotically, good prediction models should represent the true associations, however, with a finite sample, they do not necessarily represent the true underlying outcome-covariate associations, potentially leading to incorrect inferential conclusions. Given the aforementioned limitations in the current practice of association inference in the presence of many potential covariates in biomedical research, designing data-driven methods for this scenario would be useful.

This paper proposes BIEN, a novel approach to likelihood-based association inference in the presence of many potential covariates. BIEN leverages computationally pragmatic components, familiar to biomedical researchers, to suggest and score candidate covariate sets from a vast array of combinatorial possibilities, thereby offering biomedical researchers a familiar and efficient

means of analysis. BIEN was motivated by an investigation of longitudinal patterns of 37 self-reported symptoms over 3 time-points in association with future health-related quality of life (HRQoL) among 576 long-term survivors of childhood cancer. After manual engineering of many covariates (patterns/features) from longitudinal symptoms as being hypothesized to potentially influence future HRQoL, our statistical analysis was focused on identifying a small subset of covariates that are associated with two domains of HRQoL.

2.2. Methods

The goal of association inference in the presence of many potential covariates is evaluating the large set of potential covariates for their associations with the outcome of interest. We focus here on the association inference in generalized linear models, $h(E[Y|X]) = X^T\beta$, where Y is an outcome random variable whose probability distribution belongs to the exponential family, X is a covariate vector, h is a link function, and β is a parameter vector. With many potential covariates, association inference translates into (1) selecting covariates X that are deemed to be truly associated with the outcome and (2) estimating the underlying regression coefficients β that describe the associations. To achieve the goals (1) and (2), we propose BIEN and explain it through its five components. We then discuss three alternative approaches in the presence of many potential covariates.

2.2.1. BIC Elastic Net (BIEN): An Inference-focused Regression Modelling Approach

We intended to create a computationally efficient regression framework to enable likelihood-based inference, that is familiar to biomedical researchers, in the presence of many potential covariates, taking into account the relationships among them. Towards this goal, BIEN utilizes EN's penalized regression optimization function to generate candidate covariate sets (component 1: covariate set generation), MLE to estimate candidate models' regression coefficients (component 2: model estimation), BIC to score candidate models (component 3: model scoring), a truncated grid search to efficiently search for the candidate models (component 4: model search), and a backward elimination to increase the parsimony of the final selected model (component 5: model pruning). We explain these components below and describe the full algorithm of BIEN in Figure 2.1.

Component 1: Covariate set generation by EN

To suggest candidate covariate sets, BIEN uses EN, a regression method that is resistant to overfitting through an optimization function that penalizes models' goodness of fit for their

complexity. The complexity is measured in terms of sizes of regression coefficients. EN integrates both the coefficient parameters and the hyper-parameters controlling the extent and shape of the penalization in a single optimization function as follows:

$$\min_{\beta_0,\beta}\left\{\frac{1}{n}\sum_{i=1}^{n} l(y_i, \beta_0 + \sum_{j=1}^{p}\beta_j x_{ij}) + \lambda\left[\alpha\sum_{j=1}^{p}|\beta_j| + (1-\alpha)\sum_{j=1}^{p}\beta_j^2/2\right]\right\} \quad (1)$$

where n is the number of subjects, l is the negative log-likelihood, p is the number of covariates, y_i is the outcome for the ith subject, β_0 is the intercept, β_j is the regression coefficient of the jth covariate, x_{ij} is the value of the jth covariate for the ith subject, λ ($\geq$0) is the hyper-parameter determining the extent of penalization balancing the goodness of fit versus model complexity, and α ($0\leq\alpha\leq1$) is the hyper-parameter determining the shape of the penalization term through weights assigned to the sum of squared and the sum of absolute values of the regression coefficients [53]. BIEN uses EN with a given pair of (α, λ) values to generate a candidate covariate set for evaluation that consists of covariates with non-zero coefficients resulting from optimizing (1). Note that all covariates in the initial pool are simultaneously evaluated in the optimization function (1), accounting for the relationships among all covariates. This enables the selection of covariates into the model based on their joint effects, which is of specific interest for association inference in biomedical research.

Component 2: Model estimation by MLE

BIEN utilizes MLE to estimate the regression coefficients of the candidate models, which is known for its consistency under a set of regularity conditions [54, 55]: this property makes MLE particularly appealing for association inference. EN, on the other hand, is primarily focused on prediction and may yield inconsistent estimates for parameter estimation/inference [56, 57].

Component 3: Model scoring by BIC

BIEN utilizes BIC to score each of the fitted candidate models. Being negative log-likelihood values of a given model penalized for its complexity, BIC given in component 2 and EN are similar:

$$-2\sum_{i=1}^{n} l(y_i, \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}) + log\ (n) \times k \qquad (2)$$

where k is the number of parameters. Note, however, that BIC is solely a scoring method, neither indicating how to generate covariate sets from a large pool of covariates nor how to estimate regression coefficients for a covariate set, but EN does both for a given pair of (α, λ) values.

BIC has several appealing properties for association inference, in particular, model selection. First, BIC is consistent in that the true underlying model has the minimum BIC if it is among the models being considered under certain regularity conditions when the sample size approaches infinity [58]. Second, BIC approximates Bayes Factor and is consistent with Bayesian logic and interpretation [59]. Third, BIC is easy to compute as it does not require setting the priors, making it a pragmatic Bayesian approach for high-dimensional setting. Lastly, BIC can compare non-nested models [59].

Component 4: Model search by a truncated grid

To find the α and λ values that produce the model with the optimal BIC, we vary each of them over a grid of plausible values. The α value could range from 0 to 1 by the design of (1). The λ value could get any value greater than or equal to zero in theory but, in practice, the upper limit of the range is defined for the given data at hand separately for each α value and is the minimum value at which all the regression coefficients of the initial pool of covariates become zero: any λ value greater than this value also leads to intercept-only model and, thus, could be skipped [60]. Then, for a given α value, BIEN considers λ from the largest to the smallest values in the defined range sequentially and reduces computation in two ways: (1) it skips estimating MLEs for λ values

which do not change the covariate set compared to the previous set; and (2) it truncates the search for the optimal λ if c subsequent models show inferior BIC values. The hyper-parameter c is called "patience" and is set by the user. Assessing λ values in a decreasing order allows for starting the search with the simplest model as more penalization corresponds to less complexity.

Aside from the computational benefits, truncating the search prevents favouring unnecessarily complex models, where BIC could become over-optimistically too low (see Results and Figure 2.4). On the other hand, truncating the search for λ immediately after the BIC gets worse could risk identifying a local optimum as the global optimum, due to the lack of a strict monotone relationship between λ and BIC: the patience hyper-parameter is incorporated to alleviate this risk.

Component 5: Model pruning by backward elimination

Once a model has been selected based on the aforementioned four components, its parsimony is increased using a backward elimination process. This step starts with the selected MLE model and involves removing its covariates one by one until BIC score stops improving.

2.2.2. Alternative Approaches

We compare our proposed approach against two commonly used regression-based approaches in biomedical research settings working with many potential covariates, as well as a simpler version of BIEN:

- EN uses a penalized optimization function (1) with a grid search of hyper-parameter values to generate candidate models. The selection of optimal EN hyper-parameter values is typically based on cross-validation using a prediction performance measure such as Area under the Receiver Operating Characteristics Curve (AUC) for binary-outcome prediction and Mean Squared Error for continuous-outcome prediction. Note that, while EN is not intended for association inference, it is a widely used method for prediction in biomedical research with many potential covariates.
- SW starts with an intercept-only model with no covariate and sequentially adds a covariate to, or removes a covariate from, the model one by one, until selection criterion no longer favours addition or removal of a covariate. A significance-based criterion is often used as the selection criterion for SW, but in this case, we use BIC to facilitate a fair comparison with BIEN.
- BIEN-B is a simpler version of BIEN excluding the final backward step. Consequently, the model selected by BIEN-B includes at minimum the covariates selected by BIEN. BIEN-B is investigated specifically to assess the additional benefits provided by the backward elimination step in the BIEN approach.

2.2.3. Analytic Software

All analyses were performed using R version 4.2.1 (R Project for Statistical Computing). The R function “glmnet” from the package “glmnet” was used for EN, and the R base function “step” was used for SW. For BIEN and BIEN-B, we wrote R functions.

2.3. Simulation Study

We conducted a simulation study to investigate the properties of EN, SW, BIEN-B, and BIEN for association inference with many potential covariates. In line with the data-driven association inference in this study, which encompasses both covariate selection and coefficient estimation, performance of these methods was evaluated in terms of their accuracy and precision in both aspects of selection and estimation. A schematic representation of the simulation study design is provided in Supplementary Figure 2.1.

2.3.1. Simulation Design

Our simulation study borrowed the potential covariate pool of 515 covariates and their values from the case study that motivated our investigation (see Section 2.4). We experimented with six random resamples from the original sample of 576 survivors (sample size = 100, 200, 300, 500, 700 and 1000), intended to represent different scenarios with respect to power for detecting associations. Sampling with replacement was used when the new sample size was bigger than the original sample size (sample size = 700 and 1000). For each experiment, after taking the resampled covariate data, we simulated the continuous outcome of HRQoL 1000 times to give 1,000 simulated datasets based on a linear model with an intercept ($\beta_0 = 20$) and 10 specific covariates selected from the 515 covariates, hereafter called the true covariates, with coefficients of 1, 2, …, 10 and SD=10 to imitate a range of effect sizes from very small (~0.1 SD) to moderately large (~1 SD). It is important to note that the remaining 505 covariates were not involved in the outcome generation, although they may appear correlated with the outcome through their correlation with the true covariates. The error term in the model was generated using independent realizations from the identical Gaussian distribution with mean zero and standard deviation of 10, denoted as N(0,10). Since the result of this simulation is not intended to derive subject-matter knowledge

about the outcome, we simply refer to the 10 true covariates as covariates 1 to 10 and the outcome as Y in this section. The simulation process could then be summarized as follows:

$$\begin{aligned} Y = 20 &+ 1 \times covariate_1 + 2 \times covariate_2 + 3 \times covariate_3 \\ &+ 4 \times covariate_4 + 5 \times covariate_5 + 6 \times covariate_6 \\ &+ 7 \times covariate_7 + 8 \times covariate_8 + 9 \times covariate_9 \\ &+ 10 \times covariate_{10} + e \end{aligned} \quad (3)$$

where e is independently and identically distributed as N(0, 10).

We then applied EN (with 10-fold cross-validated Mean Squared Error as the selection criterion), SW (with BIC as the stopping rule/selection criterion), BIEN-B, and BIEN to each of the 1000 simulated datasets for each of the six sample sizes, using the 515 potential covariates of the case study's data. To compare the four methods, we evaluated both the selected covariate sets and estimated regression coefficients. As a descriptive measure, we investigated the numbers of selected covariates across the 1000 simulations. To formally quantify the performance, for both covariate sets and coefficient estimates, we investigated *association accuracy,* defined as proximity to the truth, and *association precision*, defined as reproducibility across the 1000 simulations [61, 62]. Note that standard evaluation metrics used in traditional inference, such as coverage probabilities of interval estimates, assess only coefficient estimates conditioned on a pre-specified set of covariates and, thus, would not be proper here as our inference includes selection of variables into the model as well: tendency for selecting more covariates into the model would result in better coverage probabilities for parameter estimates of true covariates, without penalizing a high rate of false positive selections of null-effect covariates which must be considered as part of the inferential performance. In total, we considered four measures: covariate-set accuracy, covariate-set precision, coefficient-estimate accuracy, and coefficient-estimate precision, measured as follows.

(A) *Covariate-set accuracy* can be described using the agreement of the selected covariate set with the true covariate set, measured by Jaccard Similarity Coefficient (JSC) which is defined as the ratio of the number of covariates appearing in both sets (intersection) divided by the number of covariates appearing in either set (union) [63]: the average JSC over the 1000 simulations was reported.

(B) *Covariate-set precision* can be described using the agreement in each pair of simulations between the two covariate sets selected by the pair, measured by JSC: the average JSC across all pairs of the 1000 simulations was reported.

(C) *Coefficient-estimate accuracy* was measured by the agreement between the vector of estimated coefficients for the correctly selected covariates and the vector of their true coefficient values, quantified using Intraclass Correlation Coefficient (ICC) [ICC reference]: the average ICC over the 1000 simulations was reported.

(D) *Coefficient-estimate precision* was measured by the agreement in each pair of simulations between the two vectors of estimated coefficients for the variables that were correctly selected by both simulations in the pair, quantified by ICC: the average of ICC across all pairs of the 1000 simulations was reported.

2.3.2. Simulation Results

Figure 2.2 shows box plots for the numbers of selected covariates over the 1000 simulations by the four methods in the six different sample sizes. The corresponding averages and standard deviations are reported in Supplementary Table 1. The selected number of covariates were consistently higher for EN, followed by SW, BIEN-B, and BIEN: the medians ranged from 24 to 43 for EN, from 11 to 14 for SW, 7 to 10 for BIEN-B, and 6 to 8 for BIEN. The standard deviations around the mean for the numbers of selected covariates (Figure 2.2 and Supplementary Table 2.1)

suggest notably higher variation for EN compared to the other methods: the difference was becoming less noticeable with the increase in sample size. Comparing SW with BIEN and BIEN-B, SDs suggest a preference for both BIEN and BIEN-B over SW at small sample sizes, but only BIEN, not BIEN-B, was preferred over SW at larger sample sizes. Finally, BIEN selected more consistent number of covariates compared to BIEN-B, regardless of the sample size.

Figure 2.3 shows accuracy and precision estimates for both selected covariate sets and coefficient estimates as described in Section 2.3.1 from which the following four results could be observed:

(A) Regarding the accuracy of the covariate set (Figure 2.3 A), EN was less accurate than the other three methods, (with the exception of SW at sample size = 100 which showed similar accuracy): the lower accuracy of EN was increasingly notable with the increase in sample size. Comparing SW with BIEN and BIEN-B, both BIEN and BIEN-B showed an advantage over SW in sample sizes smaller than the total number of candidate covariates (i.e., sample size = 100, 200, 300, 500), but only BIEN, not BIEN-B, attained comparable performance to SW at bigger sample sizes (sample size = 700, 1000). Finally, BIEN and BIEN-B performed similarly at sample size = 100, but with the increase in sample size, BIEN showed a slightly higher accuracy of the covariate set over BIEN-B.

(B) Regarding the precision of the covariate set (Figure 2.3 B), EN was found to be less precise than the other three methods (except for SW at sample size = 100 which showed lower precision): the poorer precision of EN was more evident with the increase in sample size. In comparison to SW, both BIEN and BIEN-B appeared superior, with the extent of superiority initially increasing with larger sample sizes that are smaller than the total number of candidate covariates (sample size = 100, 200, 300) and, then, decreasing as the sample size grew larger (sample size = 500, 700, 1000). Finally, comparing BIEN vs. BIEN-B, BIEN appeared slightly

more precise than BIEN-B with more pronounced differences with the increase in sample size. Overall, all four methods demonstrated improved precision with an increase in sample size, with notable improvements observed in SW, BIEN-B, and BIEN.

(C) Regarding the accuracy of the coefficient estimates (Figure 2.3 C), EN was notably less accurate than the other three methods, regardless of the sample size. The comparison of SW with BIEN and BIEN-B showed superiority for both BIEN and BIEN-B over SW in sample sizes smaller than or approximately equal to the total number of candidate covariates (i.e., sample size = 100, 200, 300, 500), but only BIEN, not BIEN-B, remained superior to SW at larger sample sizes (sample size = 700, 1000): all three methods had a tendency for better accuracy with increasing sample sizes. Finally, the comparison between BIEN and BIEN-B indicated slightly better accuracy for BIEN, approximately to the same extent across all sample sizes.

(D) Regarding the precision of the coefficient estimates (Figure 2.3 D), EN appeared less precise than the other three methods at all sample sizes. Comparing SW with BIEN and BIEN-B, SW appeared superior to BIEN-B at all sample sizes except 100, while BIEN appeared superior to SW at all sample sizes.

To summarize, the findings from all four performance measures (A)-(D) consistently showed that BIEN was either superior or on par with the other three methods, while EN consistently performed inferior to the others. In between these two, BIEN-B and SW were closely matched with the accuracy measures showing a preference for one over the other depending on the sample size, and the precision of covariate sets always preferring BIEN-B and the precision of coefficient estimates generally preferring SW.

The case of $\alpha = 1$, the Least Absolute Shrinkage and Selection Operator (Lasso) penalty, requires special attention. BIEN consistently showed a high tendency towards $\alpha = 1$, with 937 to 964 simulations selecting $\alpha = 1$ across all six sample sizes of the simulation study. This implies that, despite we do not have BIEN result for the specific case of $\alpha = 1$ for each single simulation, we could expect similar performance.

2.3.3. Behavior of BIC Selection in BIEN and SW

Figure 2.4 illustrates the behaviour of BIC selection for BIEN using the full grid search for the six different sample sizes. BIC values are shown for one specific simulation (simulation #1 from the 1000 simulated datasets) and one specific α value ($\alpha = 1$ from the 10 assessed α values) over the full λ grid. BIC started with a decreasing trend followed by an increasing trend at all sample sizes. At sample sizes smaller than the number of candidate covariates (i.e., sample size = 100, 200, 300), this trend was followed by a decreasing trend, and/or big jumps between small and large values as the number of covariates was getting closer to the sample size. As seen in the figure, if the search for λ values had not been truncated, BIEN would have opted for a model with many covariates for sample size = 100 and 200, instead of the model selected by the truncated search shown by the dashed line.

Figure 2.5 shows the behaviour of BIC selection for BIEN compared to SW for sample size = 100. To understand this behaviour, BIC values are shown at one specific simulation (simulation #916 from the 1000 simulated datasets) and one specific α value ($\alpha = 1$ from the 10 assessed α values), over the full λ grid for both SW and BIEN. The main x-axis shows the step number for SW, and the λ values for BIEN. While BIEN stopped after a few iterations and selected a model with a few covariates at the turning point of the BIC values, SW continued the search and selected many covariates. It is worth noting that the specific simulation #916 was selected for this figure as, at

this simulation, SW led to the highest number of selected variables across the 1000 simulations. The plot for BIEN is shown for two different patience values of 5 and 20 to illustrate the role of patience hyperparameter.

2.4. Analysis of the Case Study by BIEN

BIEN was utilized in a study aimed at inferring covariate associations with future Health-Related Quality of Life (HRQoL). The study involved a large pool of potential covariates, encompassing demographic factors, cancer-related variables, and various longitudinal symptom patterns. Using this extensive pool of covariates, BIEN inferred a covariate set and the corresponding covariate-outcome associations.

2.4.1. Study Population, Outcome Variable, and Independent Variables

The sample in the study included 576 adult survivors of childhood cancer, enrolled in both Childhood Cancer Survivor Study (CCSS) and St. Jude Lifetime Cohort Study (SJLIFE), who filled three longitudinal symptom surveys over a period of 20 years, completed at median calendar years of 1996 (range 1995-2012), 2008 (range 2007-2013), and 2013 (range 2008-2015), respectively. Briefly, both CCSS and SJLIFE are retrospectively-constructed cohorts with prospective follow-up of childhood cancer patients who survived at least 5 years post cancer diagnosis, designed for assessing late-effects of childhood cancer and its treatment. Details of the studies have been published previously [39-41].

Outcomes of interest in this case study were mental and physical component summaries (MCS and PCS, respectively) of HRQoL based on the 36-Item Short Form Survey (SF-36). Both outcomes are continuous with higher scores indicating better HRQoL, and are standardized to have mean 50 and SD 10 in the US normative population. The median calendar year for completing the HRQoL survey was 2016 (range 2010-2017), with a median of two years following the last symptom survey. The set of potential covariates included eight demographic covariates; 27 cancer-related covariates; and 480 longitudinal symptom pattern covariates. Demographic covariates included age at the baseline survey, time between baseline survey and HRQoL survey, age at the third

symptom survey, time between the third symptom survey and HRQoL survey, age at diagnosis of cancer, sex (female vs male), race (white vs non-white), and educational attainment at baseline (college graduate or higher vs. less than college). Cancer-related covariates included diagnosis group (leukemia, Hodgkin lymphoma, non-Hodgkin lymphoma, osteosarcoma, Wilms tumor, neuroblastoma, central nervous system tumors, and other diagnoses), exposure (yes/no) to each specific chemotherapeutic agent (methotrexate, intrathecal methotrexate, high dose methotrexate, cytarabine, intrathecal cytarabine, high dose cytarabine, bleomycin, alkylating agents, anthracyclines, corticosteroid, plant alkaloid, platinum agents), exposure to radiation (yes/no) at each specific site (brain, neck, chest, abdomen, pelvis), amputation (yes/no), and other surgical procedures (yes/no).

Longitudinal symptom patterns were generated based on the three symptom surveys addressing 37 questions about presence/absence of specific *symptom items*. These 37 *symptom items* were categorized into 10 *symptom sub-domains*, including anxiety (six *symptom items*), depression (six *symptom items*), sensory (eight *symptom items*), movement (four *symptom items*), cardiac (three *symptom items*), respiratory (two *symptom items*), pain (four *symptom items*), gastrointestinal (one *symptom item*), fatigue (two *symptom items*), and memory (one *symptom item*). Symptoms were further categorized into two *global domains* for psychological symptoms (anxiety and depression sub-domains) and somatic symptoms (sensory, movement, cardiac, respiratory, pain, gastrointestinal, fatigue, and memory sub-domains). Cross-sectional symptom summary variables were generated at each time-point by counting the number of present symptom items overall, eight individual sub-domains containing more than one items, and two global domains capturing psychosocial and somatic symptoms. With these, we had 48 cross-sectional symptom measures including 37 *symptom items* and 11 symptom summary variables.

2.4.2. Longitudinal Symptom Patterns

For each cross-sectional symptom measures above, we constructed 10 longitudinal symptom patterns hypothesized to affect future HRQoL. These included six consistent presence or absence patterns and four increase or decrease patterns. A consistent presence (absence) pattern is when the symptom is consistently present (absent) at a pair of successive time-points or at all three time-points, resulting in six patterns (presence/absence × three sets of time-points). Specifically, they include: (1) Consistent Presence at T1 & T2 but not T3; (2) Consistent Presence at T2 & T3 but not T1; (3) Consistent Presence at T1, T2, & T3; and (4)-(6) replacing Presence with Absence in (1)-(3). An increase (decrease) pattern indicates an absolute increase (decrease) of the symptom from an earlier time-point to its successive time-point. Increase/decrease patterns were constructed for each pair of successive time-points, resulting in four patterns (increase/decrease × two sets of time-points). Specifically, these patterns include: (1) Increase from T1 to T2; (2) Increase from T2 to T3; (3) Decrease from T1 to T2; and (4) Decrease from T2 to T3. With these definitions, we had 480 longitudinal symptom patterns (10 patterns × 48 cross-sectional symptom measures). Of note, for symptom summaries, the consistency patterns were concerned with any presence (i.e., count>0), while increase/decrease patterns were concerned with counts of present symptom items.

2.4.3. Association Inference by BIEN

We then applied BIEN for the set of 515 potential covariates for the outcomes of MCS and PCS. Table 2.1 reports the covariates associated with the two HRQoL outcomes, their coefficient estimates and the corresponding 95% confidence intervals and p-values. Selected model for MCS included only symptom patterns with no demographic and cancer-related covariates, indicating that symptom patterns may mediate the influence of these other variables on the outcomes. The selected model for PCS included age at the 3rd symptom survey along with symptom patterns.

Regarding the type of patterns, consistent absence at three time-points was the most frequently selected pattern in association with both HRQoL outcomes (consisting of four out of the eight patterns for MCS, and four out of the five patterns for PCS), and was always associated with higher HRQoL. Consistent presence at all three time-points was the next frequent pattern (appearing two times for MCS and once for PCS), and was always associated with lower HRQoL. The model for MCS also included one consistent presence pattern over two time-points, associated with lower HRQOL. Besides the consistency patterns, the model for MCS also included one increase pattern from T2 to T3, associated with lower MCS.

Regarding the symptom content, both psychological and somatic symptoms were associated with MCS, while only somatic symptoms were associated with PCS. Specifically with respect to the symptom sub-domains, symptoms of depression and anxiety were associated with MCS, symptoms of movement problems, cardiac and pain were associated with PCS, symptoms of fatigue were associated with both outcomes, and, finally, sensory, respiratory and gastrointestinal symptoms were not identified to be associated with either of the outcomes.

2.5. Discussion

We considered here the problem of drawing inference on covariate-outcome associations in the presence of many potential covariates. With many potential covariates, covariate selection into the model is challenging, especially when covariates are correlated, and must be protected from overfitting. Hypothesis testing and estimation of parameters of covariates that are truly associated with the outcome need to account for the relationship among the potential covariates and must be simultaneously considered in a model. To overcome the challenges associated with conducting the standard likelihood-based inference via MLE in the presence of many potential covariates, we proposed BIEN. BIEN utilizes familiar and widely used concepts of EN, BIC, a truncated grid search, and backward elimination, each specifically selected to address different challenges in a practical and computationally manageable manner. By leveraging EN, BIEN could consider all covariates in a large set simultaneously and suggest a series of candidate covariate sets effectively according to the values of its two hyper-parameters, α and λ. Finer grids of these hyper-parameters would increase BIEN's chance of finding the true underlying model but add computational burden: by utilizing a truncated grid search, BIEN explores a limited range of hyperparameters while aiming to maintain good performance. BIC is employed to protect against overfitting and, finally, backward elimination helps avoid including redundant variables in the final covariate set. All these elements of BIEN are conductible using the standard statistical software and familiar to biomedical researchers. Our simulation experiment, based on the real dataset, showed notable advantages of BIEN and BIEN-B over EN for inference with many potential covariates, but found performance preference between BIEN-B and SW with BIC to vary depending on sample size, and finally, consistently showed similar or superior performance for BIEN over the other methods.

Unlike BIEN for which the variation of the number of selected covariates across the 1000 simulations appeared similar across all sample sizes, the decrease in sample size led to the over-selection of covariates by SW, contrary to our prior expectation that BIC would protect both SW and BIEN against overfitting, specifically at sample sizes smaller than the number of candidate covariates (See Figure 2.2 A). This must be attributable to the difference in the candidate covariate-set generation between SW and BIEN. While SW's use of BIC employs the penalization concept in model selection stage, BIEN additionally uses the penalization concept in generating candidate covariate sets through EN's optimization function (1). For SW, at each step, all possible additions and removals of a covariate to/from the current covariate set are considered with re-estimation of all covariates' regression coefficients in the model. With small sample sizes, this highly flexible process can overfit. For BIEN, on the other hand, covariate sets are suggested by EN's penalized optimization function solved over the range of penalty values. This reduced flexibility lowers the chance of overfitting to the observed data. It is worth noting that even in the extreme cases for which SW led to selecting many covariates (with continuously-decreasing BICs up to the end), the trend of BIC values in BIEN with the truncated grid search remained as expected (with a decreasing followed by an increasing trend) (See Figure 2.5 for one example). This distinguished the inferential performance of BIEN compared to SW with small sample sizes. Deviations from the expected patterns in both SW and BIEN with the full grid search, in scenarios with the number of candidate covariates greater than the sample size, could suggest against using a highly flexible statistical method for such cases. Aside from the observed overfitting and the unexpected behaviour of BIC for SW in small sample sizes, the computations for SW could become considerably high with the number of potential covariates, while BIEN is computationally much less demanding using EN in its design.

Penalized or regularized regression (e.g., EN and Lasso) techniques, despite not being initially proposed for association inference, have been adopted in practice, sometimes in combination with additional steps, for selecting covariates for an MLE model, due to their efficient search of the covariate space, especially when the number of candidate covariates is large. The most straightforward approach seems to be fitting an MLE model using the selected covariates by a penalized/regularized regression method. However, our EN simulation results showed that EN or even Lasso, a more stringent approach with more regression coefficients being driven to zero, often selected too many covariates, indicating poor *covariate-set accuracy* and overfitting in the subsequent MLE model. This is because the best covariate set is selected based on the performance of the penalized models, which may not necessarily reflect the performance of corresponding MLE models. The original penalized model might have controlled overfitting by attenuating coefficient estimates despite the inclusion of many covariates, but re-estimating these coefficients with MLE often lead to overfitting. A more involved approach is to fit MLEs for various covariate sets obtained from the penalized/regularized regression across the range of penalty hyperparameter values and selecting based on the performance of these subsequent MLEs. This two-step process was originally introduced to improve prediction performance, but could also mitigate issues regarding association inference such as the over-selection of covariates, the high bias in coefficient estimates introduced by shrinkage, and the inconsistency of the estimates. While this approach shares similarities with the first two components of the BIEN framework, it is typically more liberal in the selection of covariates due to relying on cross-validation of prediction performance rather than BIC and not utilizing backward elimination. Extensions of Lasso/EN including the aforementioned methods that attempt to loosen the tight control of coefficient estimation by the

same hyperparameter controlling covariate selection are often collectively referred to as Relaxed Lasso/EN [64].

The truncated grid search was introduced to BIEN after observing the complications corresponding to a full grid search approach that is initially employed for selecting hyper-parameters of EN, which is standard in the application of EN. Figure 2.4 suggests that replacing the full grid search with the truncated grid search helps BIEN avoid overfitting for scenarios with smaller sample sizes where BIC behaves poorly with the number of covariates getting close to the sample size. The initial decreasing values of BIC followed by increasing values is expected as we anticipate BIC to improve with true covariates entering the model and to worsen with false-positive covariates entering the model afterwards. The subsequent emerging deviations from this pattern (decreasing pattern and sudden jumps) indicates improper behaviors of likelihood and BIC with too many parameters relative to the sample size. Figure 2.1 suggests that the truncated grid search with the patience parameter could help finding the turning point of BIC before arriving at such problematic BIC values.

A difficulty of BIEN for an end user could be the requirement of the hyper-parameter grids to be set by the user. In both our simulation and case study, we used the same hyper-parameter grids and a default patience parameter. In practice, the appropriateness of these values can be ascertained by further investigations as explained below. Specifically, for α hyper-parameter, we initialized the candidate grid to include 10 values of 0.1, 0.2, 0.3, …, and 1.0. Due to the large set of potential covariates, we did not investigate $\alpha = 0$ as it would retain all covariates in the set which is not a realistic covariate set. The selected covariate set in our simulation almost always corresponded to $\alpha = 1$ (i.e., Lasso), however. While $\alpha = 1$ is not always optimal in general, confining to $\alpha = 1$ instead of a candidate grid would reduce the computations of BIEN approximately by a factor of

10. For λ hyper-parameter, we initialized the candidate grid to include 1000 values, ranging from 0 to the minimum value at which all the regression coefficients of the initial pool of covariates become zero, using a uniform distribution on the natural logarithmic scale. We used the 1000 values in the λ grid to make it sufficiently dense to allow suggesting models of various sizes. Furthermore, we used the logarithmic scale for λ grid with the aim that the sizes of candidate covariate sets corresponding to the λ grid are more equally spaced: a uniformly distributed λ grid leads to bigger jumps in larger sizes of candidate covariate sets. Since λ hyper-parameter determines the extent of shrinkage, a careful set up of an appropriate candidate grid is essential for BIEN's success, and, thus, users are encouraged to examine the size of the candidate sets for some variations of the λ grid. For patience hyper-parameter, we used a default value of 5 for all our implementations. To assess the appropriateness of this value, users are encouraged to draw plots similar to Figure 2.4 and Figure 2.5, and make sure that this value is big enough to avoid selecting the local minimums of BIC.

A key consideration when interpreting the result of BIEN for epidemiological association inference is its ineffectiveness, common to other high-dimensional variable selection tools, to automatically disentangle the influence of confounding factors. Although the statistical literature is well developed for controlling for confounding in traditional analyses, where the interest lies in assessing the effect of a single exposure adjusting for *a priori* hypothesized potential confounders, confounding investigation is beyond the scope of this work due to the lack of a single exposure of interest and specific hypotheses around it. Henceforth, interpreting suggested covariate-outcome associations derived from BIEN requires further analysis and consideration of confounding factors after obtaining results from the BIEN analysis.

Our case study shows the appearance of various symptom representations including both consistent presence/absence and increase patterns, both two and three time-point-based patterns, and both individual symptom items and symptom summaries which highlights the importance of designing methods for association inference that can work with many potential covariates such as the pool of covariates in the case study rather than focusing only on a handful of pre-defined covariates.

Our simulation investigation borrows the set of potential covariates from a real case study where prevalence of covariates and correlation between covariates were designed to reflect the complexities of real data. In our investigation, we exclusively focused on continuous outcome, and anticipate a similar performance for binary, count, and time-to-event outcomes, which needs to be investigated in future work. We hope that our proposed data-driven method and evaluation framework will encourage development of additional methodologies and assessment measures in the context of association inference in the setting of many potential covariates and provide insight to scientific questions regarding true underlying models.

Table 2.1: Selected model by BIEN for Mental and Physical Component Score outcomes.

Mental Component Score Outcome				
Covariate Name (*Symptom Sub-Domain:* Symptom Item or **Summary of Items** or ***Symptom Global Domain*: Summary of Items**)	**Pattern Type**	**Estimate**	**95% CI**	**P-value**
(Intercept)		40.65	38.09, 43.22	<0.001
Depression: Feeling no interest in things	Consistent Absence at T1, T2, & T3	5.00	2.97, 7.03	<0.001
Depression: Feeling hopeless about the future	Increase from T2 to T3	-7.09	-10.06, -4.11	<0.001
Depression: **Summary of Items ***	Consistent Presence at T1, T2, & T3	-5.14	-8.35, -1.94	0.002
Anxiety: Suddenly scared for no reason	Consistent Presence at T1 & T2 but not T3	-14.07	-21.15, -7.00	<0.001
Anxiety: Feeling tense or keyed up	Consistent Absence at T1, T2, & T3	2.85	1.02, 4.68	0.002
Anxiety: So restless cannot sit still	Consistent Absence at T1, T2, & T3	3.00	0.96, 5.04	0.004
Fatigue: Feeling weak	Consistent Absence at T1, T2, & T3	2.84	0.85, 4.84	0.005
***Somatic*: Summary of Items ***	Consistent Presence at T1, T2, & T3	-2.27	-3.92, -0.61	0.007
Physical Component Score Outcome				
Covariate Name (*Symptom Sub-Domain:* Symptom Item or **Summary of Items** or ***Symptom Global Domain*: Summary of Items**)	**Pattern Type**	**Estimate**	**95% CI**	**P-value**
(Intercept)		46.10	41.06, 51.14	<0.001
Demographic: Age at 3rd Time-point [Year]		-0.33	-0.43, -0.23	<0.001
Movement: Weakness/inability to move leg	Consistent Absence at T1, T2, & T3	5.82	3.12, 8.52	<0.001
Cardiac: Chest pain with exercise	Consistent Absence at T1, T2, & T3	5.41	3.54, 7.28	<0.001
Pain: Prolonged pain in arms, legs, or back	Consistent Absence at T1, T2, & T3	2.91	1.11, 4.71	0.002
Fatigue: Feeling weak	Consistent Absence at T1, T2, & T3	6.01	4.07, 7.95	<0.001
***Somatic*: Summary of Items**	Consistent Presence at T1, T2, & T3	-2.37	-4.04, -0.70	0.006

Note: * Counting Present Symptom Items

Algorithm of BIEN

Create candidate values for α hyper-parameter of EN ranging from zero to one

for each value of α **do**
Fit an intercept-only MLE model with no covariate
Initialize the optimal BIC to be BIC of the fitted intercept-only MLE model
Initialize the covariate set to be empty
Initialize the patience counter to zero
Find the minimum value of λ hyper-parameter of EN at which all the regression coefficients of the initial pool of covariates become zero
Create candidate values for λ ranging from zero to the minimum found above

for each value of λ in a descending order **do**
if the patience counter = c* **then** break the **for** loop
use EN with the current α and λ and define the new covariate set to be the covariates with non-zero coefficients identified by EN
if the new covariate set differs from the current covariate set **then**
estimate the regression coefficients of the new covariate set by MLE
if BIC of the fitted MLE model is better than the current optimal BIC **then**
Update the optimal BIC
Set the patience counter to zero
else
Add one to the patience counter
end if
end if
end for
end for

Initialize the optimal covariate set with the covariate set of the model with the optimal BIC above
repeat
if the optimal covariate set is empty **then** break the **repeat** loop
Let n be the number of the covariates in the optimal covariate set
Fit n MLE models, each excluding one covariate from the optimal covariate set
Find the model with the lowest BIC among the n models
if BIC of the new MLE model is better than the current optimal model's BIC **then**
Update the optimal BIC
Update the optimal covariate set
else
break the **repeat** loop
end if
end do

Return the covariate set and MLE coefficient estimates of the model with the optimal BIC

Figure 2.1: Algorithm of BIEN.

Note: *c = Maximum number of successive evaluations of no BIC improvement before breaking the λ loop; BIEN: Bayesian Information-Criterion Elastic Net; BIC: Bayesian Information Criterion; EN: Elastic Net; MLE: Maximum Likelihood Estimate.

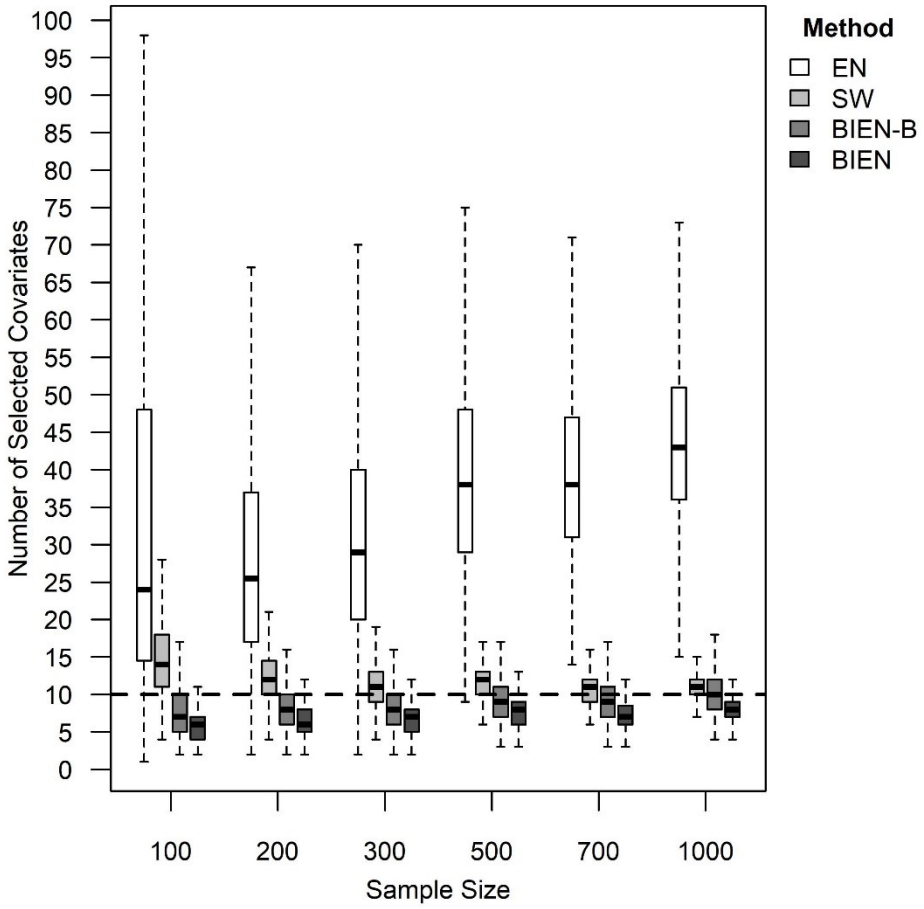


Figure 2.2: Box plots of the number of selected covariates in the simulation experiments.

Box plots correspond to the four methods of EN, SW, BIEN-B, and BIEN with six different sample size scenarios (sample size = 100, 200, 300, 500, 700, 1000) over 1000 simulations.

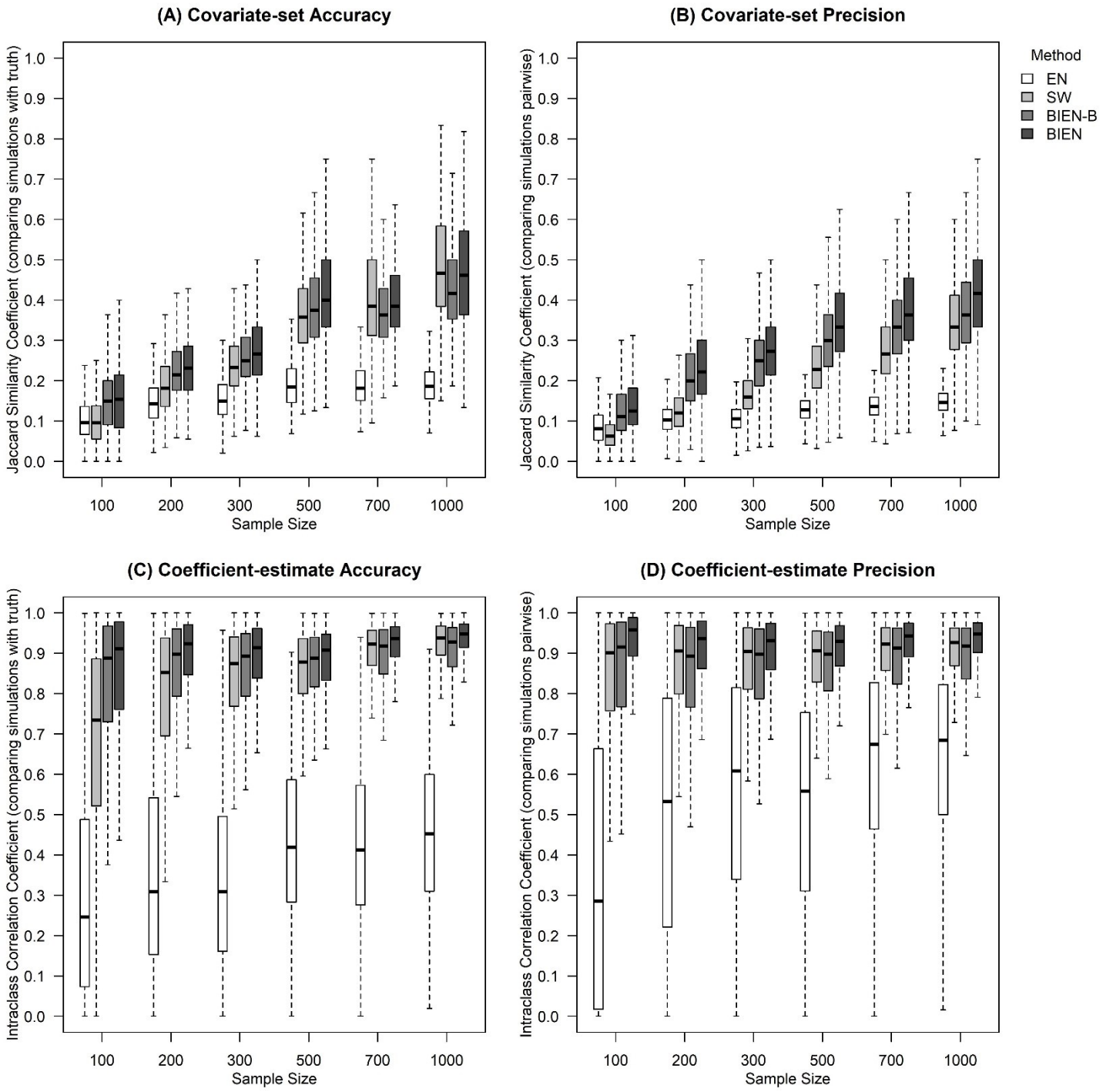


Figure 2.3: Box plots of the performance of methods in the simulation experiments.

Box plots of: (A) accuracy of the selected covariate set with respect to the truly-associated covariate set; (B) precision of selected covariate sets estimated from all pairs of simulation; (C) accuracy of the estimated coefficients with respect to the true coefficients among the correctly selected covariates; and (D) precision of the estimated coefficients estimated from all pairs of simulation, shown for EN, SW, BIEN-B, and BIEN with six different sample size scenarios (sample size = 100, 200, 300, 500, 700, 1000) over 1000 simulations.

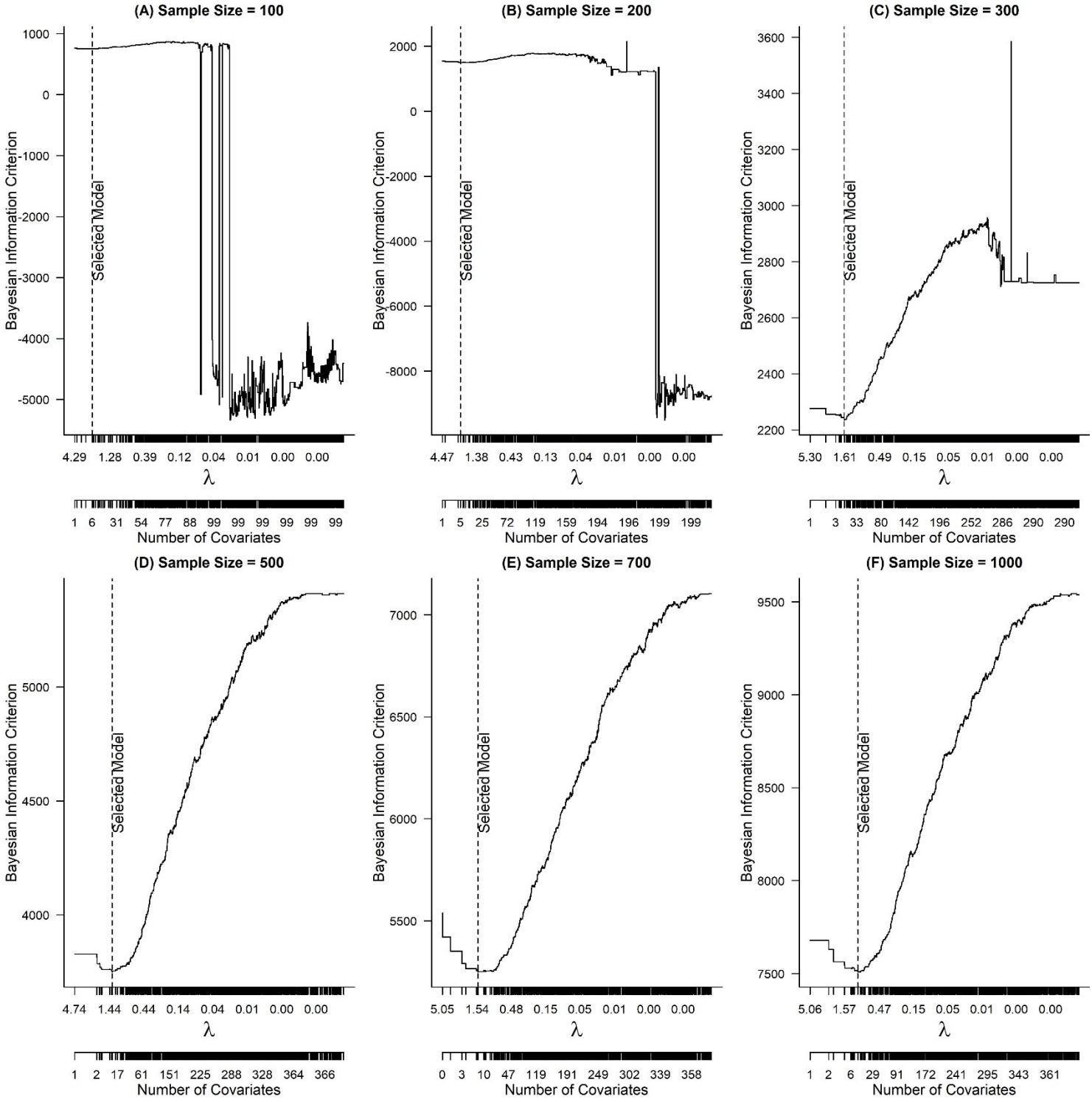


Figure 2.4: BIEN with full grid search exploration.

Results corresponding to candidate models for BIEN with the full grid search for a specific simulation (simulation #1) for the six different sample sizes (A) sample size = 100, (B) sample size = 200, (C) sample size = 300, (D) sample size = 500, (E) sample size = 700 and (F) sample size = 1000. Plots are drawn for a specific α value ($\alpha = 1$), with the x-axis showing the λ values in the descending order to represent the computation order and the y-axis showing the corresponding BIC values. The λ axis is drawn in a logarithmic scale. The corresponding number of selected covariates is shown in a second x-axis. The dashed line corresponds to the selected model by BIEN using the truncated grid search.

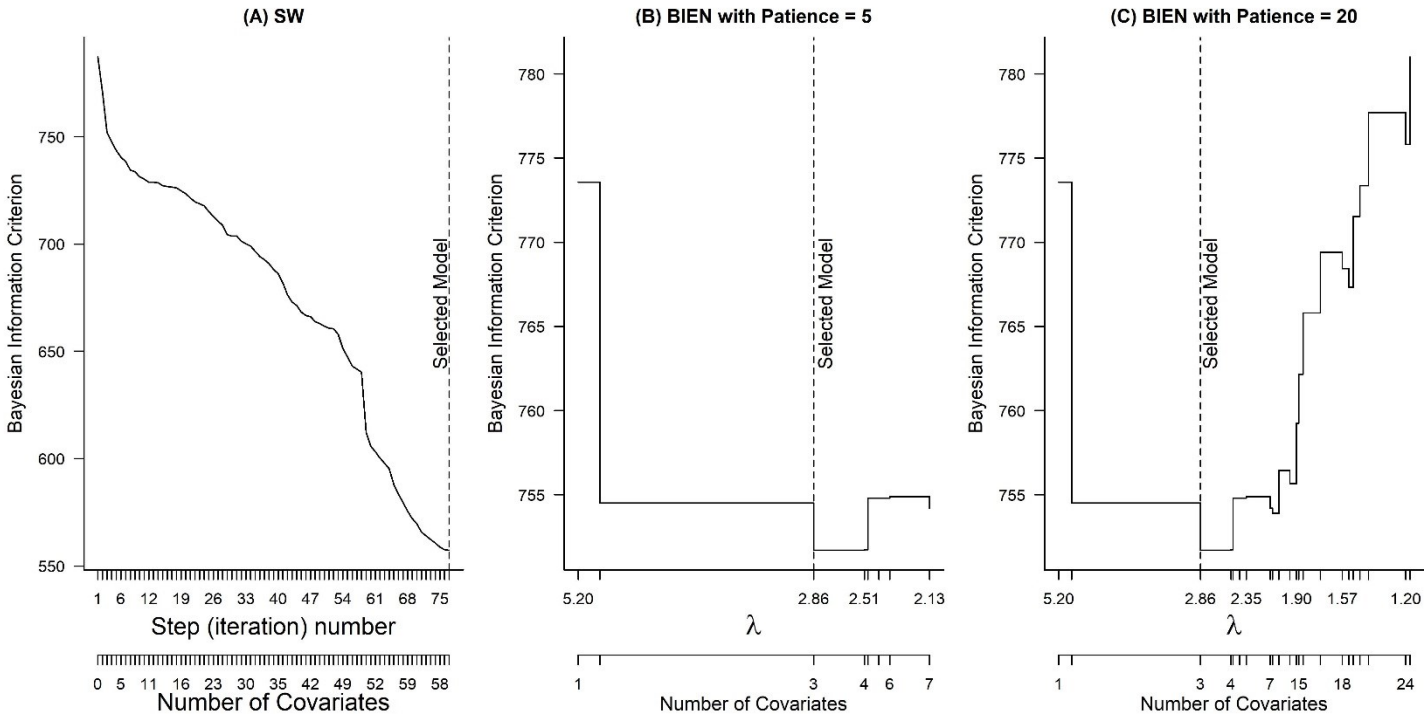


Figure 2.5: Method performance exploration under small sample size scenario.

Results corresponding to candidate models for a specific simulation (simulation #916) for sample size = 100 for (A) SW and (B) BIEN with the truncated grid search with patience of 5 and (C) BIEN with the truncated grid search with patience of 20. The plot corresponding to SW shows the step number as the x-axis and the corresponding BIC values as the y-axis. The plots corresponding to BIEN is drawn for a specific α value ($\alpha = 1$), with the x-axis showing the λ values in the descending order to represent the computation order and the y-axis showing the corresponding BIC values. The λ axis is drawn in a logarithmic scale. For both plots, the corresponding number of selected covariates is shown in a second x-axis. The dashed lines correspond to the selected models.

2.6. Supplementary Information

Supplementary Table 2.1. The average numbers of selected covariates (standard deviations) by EN, SW, BIEN-B, and BIEN with the six different sample size scenarios over 1000 simulations.

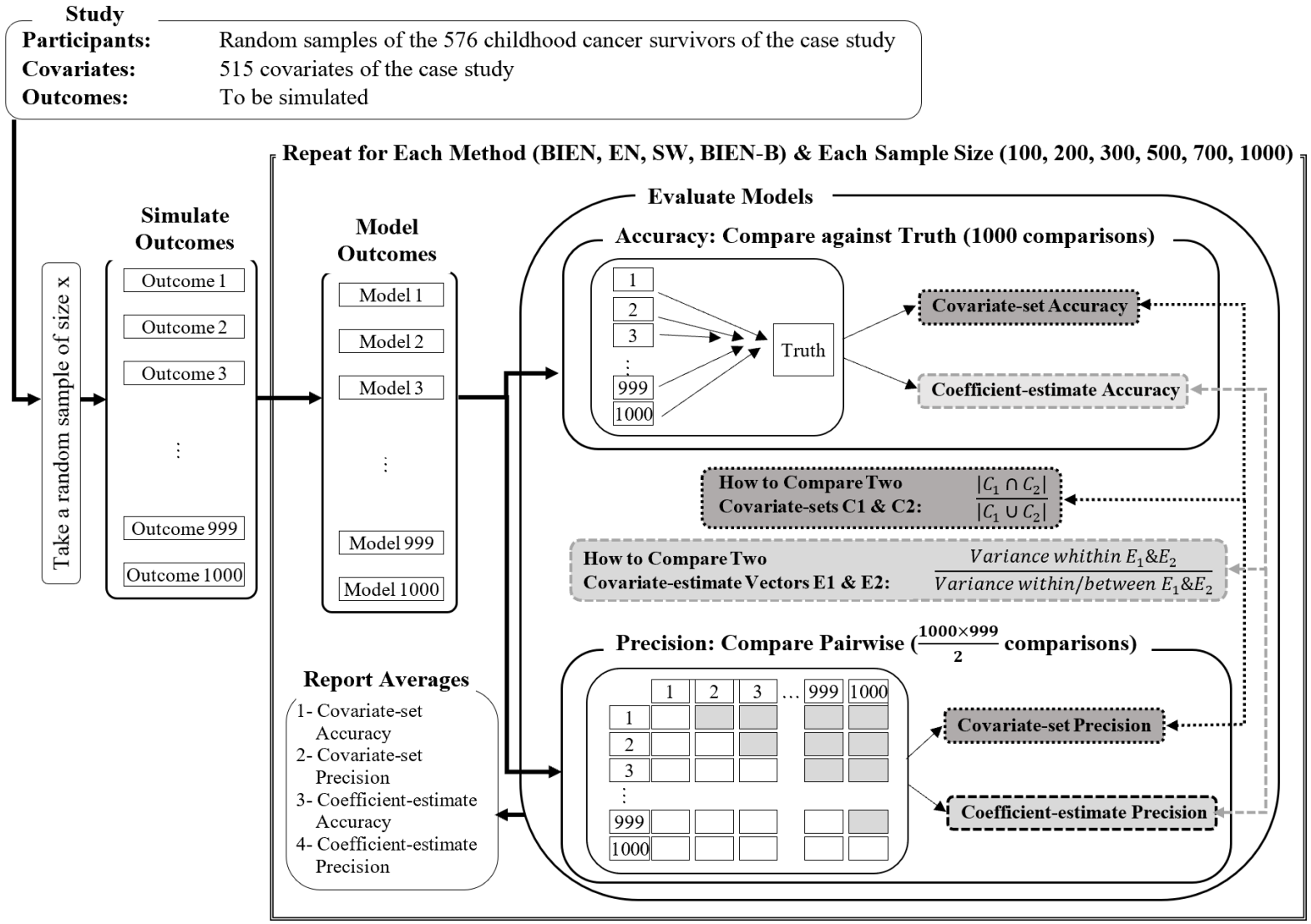
Number of Selected Covariates				
Sample Size	**EN***	**SW†**	**BIEN-B‡**	**BIEN§**
100	53.8 (90.2)¶	15.2 (6.7)	7.7 (3.4)	6.0 (2.2)
200	31.7 (35.6)	12.5 (4.0)	8.1 (3.0)	6.5 (2.0)
300	33.5 (24.7)	11.6 (3.1)	8.1 (2.9)	6.7 (2.0)
500	39.6 (15.1)	11.9 (2.6)	9.3 (2.8)	7.7 (1.9)
700	39.8 (12.4)	10.8 (2.0)	8.9 (2.9)	7.3 (1.8)
1000	44.6 (12.1)	11.0 (1.9)	9.8 (2.9)	7.9 (1.7)

Note: * EN: Elastic Net; † SW: Stepwise selection; ‡ BIEN-B: BIEN minus the Backward step; § BIEN: Bayesian Information-Criterion Elastic Net; ¶ mean (standard deviation)

Supplementary Table 2.2. Average agreement of the selected covariates and estimated coefficients between the simulation results and the truth (accuracy) and between the result of all pairs of simulation (precision) for EN, SW, BIEN-B, and BIEN with the six different sample size scenarios over the 1000 simulations.

Covariate Sets								
	EN*		**SW†**		**BIEN-B‡**		**BIEN§**	
Sample Size	**Accuracy (JSC_{true}¶)**	**Precision (JSC_{pair}‖)**	**Accuracy (JSC_{true}¶)**	**Precision (JSC_{pair}‖)**	**Accuracy (JSC_{true}¶)**	**Precision (JSC_{pair}‖)**	**Accuracy (JSC_{true}¶)**	**Precision (JSC_{pair}‖)**
100	0.10	0.09	0.11	0.07	0.15	0.13	0.16	0.14
200	0.15	0.11	0.19	0.13	0.23	0.21	0.24	0.24
300	0.16	0.11	0.24	0.17	0.26	0.26	0.27	0.29
500	0.19	0.13	0.37	0.24	0.38	0.31	0.41	0.35
700	0.19	0.14	0.41	0.28	0.38	0.35	0.41	0.39
1000	0.19	0.15	0.48	0.35	0.42	0.38	0.47	0.43
Coefficient Estimates								
	EN†		**SW‡**		**BIEN-B**		**BIEN***	
Sample Size	**Accuracy (ICC_{true}**)**	**Precision (ICC_{pair}††)**	**Accuracy (ICC_{true}**)**	**Precision (ICC_{pair}††)**	**Accuracy (ICC_{true}**)**	**Precision (ICC_{pair}††)**	**Accuracy (JSC_{true}§)**	**Precision (JSC_{pair}¶)**
100	0.31	0.37	0.68	0.83	0.80	0.82	0.84	0.93
200	0.35	0.50	0.79	0.86	0.84	0.83	0.88	0.91
300	0.34	0.56	0.83	0.87	0.85	0.85	0.88	0.90
500	0.43	0.52	0.85	0.88	0.86	0.86	0.88	0.91
700	0.42	0.63	0.90	0.90	0.88	0.87	0.92	0.92
1000	0.46	0.64	0.92	0.90	0.90	0.88	0.94	0.93

Note: * EN: Elastic Net; † SW: Stepwise selection; ‡ BIEN-B: BIEN minus the Backward step; § BIEN: Bayesian Information-Criterion Elastic Net; † EN: Elastic Net; ‡ SW: Stepwise selection; ¶ Jaccard Similarity Coefficient of the selected covariate set with the true covariate set; || Jaccard Similarity Coefficient in each pair of simulations between the two covariate sets selected by the pair; ** Intraclass Correlation Coefficient of the vector of estimated coefficients for the correctly selected covariates and the vector of their true coefficient values; †† Intraclass Correlation Coefficient in each pair of simulations between the two vectors of estimated coefficients for the variables that were correctly selected by both simulations in the pair.



Supplementary Figure 2.1. Schematic representation of the simulation study design.

CHAPTER 3

Longitudinal Patient-Reported Symptom Patterns for Modelling Future Health-Related Quality of Life in Childhood Cancer Survivors: A Machine Learning Approach

3.1. Introduction

Advances in the treatment of childhood cancer have increased the five-year survival rates substantially, from 20% in 1950-1954 to more than 85% currently in the US [4, 65, 66]. With this success comes the growing concern of long-term adverse effects of cancer and its treatment, known as late-effects, in the aging population of survivors/patients [67-69]. Early identification of modifiable risk factors associated with these late-effects could help improve the lifelong well-being of survivors through timely targeted interventions [70-72]. To this end, utilizing survivors` patient-reported measures, including symptoms and health-related quality of life (HRQoL) collected directly from the survivors themselves, could provide valuable insights for delivering patient-centered survivorship care.

Patient-reported Outcomes (PROs) are measures obtained directly from the patients themselves, without the involvement of a second person, and thus, reflect the patients' own perception of their health status [73]. In cancer survivorship research, PROs could inform survivors` unique experience of cancer, encompassing their coping mechanisms and personal priorities, which are known exclusively to each survivor, and thus, might offer useful information beyond and preceding to the clinical data (i.e., before a problem becoming clinically apparent) [28, 29, 33, 67, 74]. A growing body of evidence suggests that regular collection of PROs from cancer survivors resulted in better quality of life (QoL) after six months, reduced emergency room visits after 36

months, and improved survival rates between one and eight years compared to standard care without PROs collection [30-32]. The authors of these reports attributed this improvement to earlier identification of symptoms suggesting adverse events. This, in turn, prompts discussions between patients and healthcare providers, facilitating timely intervention to potentially avert adverse consequences, including symptom management counseling, supportive medications, treatment adjustments, and referrals. Despite this widespread recognition of the importance of PROs [33], PRO collection often remains optional in clinical trials, survivorship or primary care visits, and statistical analysis [14, 34, 35].

HRQoL, defined as an aspect of quality of life (QoL) affected by disease or treatments [74], is a recognized multi-dimensional measure of health and well-being that could be collected directly from the patients. HRQoL has demonstrated prognostic value for subsequent development of clinical diseases and survival in cancer patients [75, 76] and its potential use as outcome in guiding clinical decision-making for cancer patients has been suggested [77].

Symptoms, defined as the subjective experience of a potential health issue, serve as importance modifiable intervention targets that could be collected directly from the patients. Symptoms have established importance as associative and predictive factors of health outcomes [7], and when identified and treated early, have shown the potential to mitigate the late-effects and reduce the associated decline in quality of life among cancer survivors [68]. Previous research utilizing symptom data has often focused on specific a priori hypothesized set of symptoms or symptom burden scores at single time-point assessments, which may limit the potential to extract novel and timely knowledge/insight through a holistic view to cancer survivors, considering their diverse and dynamic experience of risk factors experienced throughout the prolonged period of cancer survivorship continuum [14, 34, 35, 78, 79].

Cutting-edge machine learning techniques have demonstrated considerable promise in predicting outcomes in the presence of many predictors/covariates [46]. The potential of machine learning lies in its capacity to uncover patterns between covariates and outcome solely based on data, alleviating the need for explicit definition of models based on a priori information.

This study aims to uncover novel insights that could potentially be utilized to improve the long-term well-being of childhood cancer survivors. By utilizing the longitudinal data collected from the Childhood Cancer Survivor Study (CCSS) [80] and St. Jude Lifetime Cohort Study (SJLIFE) [81] longitudinally over a survivorship period and employing machine learning techniques, this investigation examines associations of numerous longitudinal patterns of 37 patient-reported symptoms over three time-points with future HRQoL in adult survivors of childhood cancer, infer a subset of associative patterns, and estimate their HRQoL associations.

3.2. Materials and Methods

3.2.1. Study Population

This study included 576 five-year survivors of childhood cancer who participated in both CCSS and SJLIFE. To be eligible, subjects had to be at least 18 years of age or older at the study baseline, have been diagnosed with cancer before the age of 21, have patient-reported (rather than proxy/caregiver-reported) symptom data collected at three time-points by CCSS/SJLIFE questionnaires, and have patient-reported HRQoL data collected subsequently by SJLIFE questionnaires.

CCSS, initiated in 1994, is a retrospectively-constructed cohort with prospective follow-up of 24,368 five-year survivors diagnosed between 1970 and 1999 who had been treated for childhood cancer at 31 institutions in North America. SJLIFE, initiated in 2007, is a retrospectively-constructed cohort with prospective follow-up of 4,094 five-year survivors diagnosed after 1962 who had been treated at the St. Jude Children's Research Hospital. Both studies assess occurrences of late-effects of childhood cancer and its treatment with prospective follow-up through survivorship in adulthood. Details of the studies have been published elsewhere [39-41].

3.2.2. Measurement

HRQoL Outcomes

HRQoL was measured using the 36-Item Short Form Survey (SF-36) [82], a widely accepted and validated questionnaire which provides eight scores representing different dimensions of HRQoL: mental health, emotional role limitation, social functioning, vitality, general health perceptions, physical role limitation, physical functioning, and bodily pain. These scores are weighted sum of the questions across different sections of the questionnaire. Additionally, two scores of mental component summary (MCS) and physical component summary (PCS) were calculated based on

the eight original scores to represent the overall mental and physical well-being [83]. All ten scores were transformed to have an average of 50 and a standard deviation of 10 in a normative population, adjusting for age and sex [84]. Higher scores indicate better health status, with a cut-off of 40 used to distinguish suboptimal HRQoL from optimal HRQoL. The median calendar year for completing the HRQoL survey was 2016 (range: 2010 to 2017) (Supplementary Figure 3.1).

Clinical Data

The clinical data utilized in this study encompassed a range of demographic, cancer-diagnosis, and cancer-treatment factors. The demographic data included age at the baseline survey, time between baseline and outcome survey, age at the 3rd symptom survey (i.e., the last survey), time between the 3rd symptom and outcome survey, age at diagnosis of cancer, sex (female vs male), race (white vs non-white), educational attainment at baseline (college graduate or higher vs less than college).

Regarding cancer diagnosis, the study considered the specific type of initial diagnosis, which included leukemia, Hodgkin lymphoma, non-Hodgkin lymphoma, osteosarcoma, Wilms tumor, neuroblastoma, central nervous system tumors, and other malignancy. For cancer treatment, the study included exposure to chemotherapy (yes/no) for each specific agent (methotrexate, intrathecal methotrexate, high dose methotrexate, cytarabine, intrathecal cytarabine, high dose cytarabine, bleomycin, alkylating agent, anthracycline, corticosteroid, plant alkaloid, platinum), exposure to radiation (yes/no) at each specific site (brain, neck, chest, abdomen, pelvis), amputation (yes/no), and other surgery (yes/no) within the first five years of primary cancer diagnosis.

Symptom Data

The symptom data utilized in this study consisted of responses to 37 questions, each assessing the presence/absence of a specific symptom, referred to as *symptom items*. These were available through comprehensive questionnaires administered to the two cohorts under investigation. The survivors in this study completed three symptom surveys over a span of 20 years, with the median calendar years (ranges) for their completion being 1996 (range: 1995 to 2012), 2008 (range: 2007 to 2013), and 2013 (range: 2008 to 2015) (Supplementary Figure 3.1).

3.2.3. Statistical Analysis

Recognizing the dynamic nature of symptoms through survivorship journey, our analytic approach initially generates various patterns of symptom existence through time as more stable characteristic of survivors, and, subsequently, employs a data-driven selection of these patterns to facilitate identification of a combinatorial set from the numerous possibilities. More specifically, the analysis framework in this work consisted of four steps. We initiated with data pre-processing (2.3.1), followed by engineering longitudinal symptom patterns (2.3.2). These patterns were then used in the modeling procedure (2.3.3), and finally, the models generated were evaluated (2.3.4). The details of these steps are explained below and a visual representation is shown in Supplementary Figure 3.2.

Preprocessing

First, any missing symptom response at each time-point, which accounted for less than approximately 2 percent of the symptom data, was replaced by its value from the preceding time-point, and if unavailable, from its successive time-point. However, if a survivor was missing responses to a symptom question at all three time-points, it was assumed that they did not experience the symptom.

Feature Engineering: Expert-hypothesized Longitudinal Symptom Patterns

To generate the longitudinal patterns for subsequent modelling, we first generate cross-sectional symptom summaries from the 37 survey-collected *symptom items* at various levels. Next, we develop longitudinal symptom patterns utilizing both the survey-collected symptoms and the generated symptom measures. These two steps are described in more details in section (A) and (B), respectively, with a concise visual representation available in Supplementary Figure 3.3:

(A) To generate cross-sectional symptom summaries, the 37 *symptom items* were first arranged into clinically meaningful domains in two steps. In the first step, *symptom items* were arranged into 10 *sub-domains*, including depression (thoughts of ending life, feeling lonely, feeling blue, feeling no interest in things, feeling hopeless about the future, and feelings of worthlessness), anxiety (nervousness or shaking inside, suddenly scared for no reason, feeling fearful, feeling tense or keyed up, spells of terror or panic, and so restless cannot sit still), sensory (decreased sense of touch, tinnitus/ringing in ear, dizziness, double vision, other trouble seeing, very dry eyes, abnormal sense of taste, and numbness), movement (problem with balance, tremors/movement problems, weakness/inability to move arm, and weakness/inability to move leg), cardiac (arrhythmia, angina pectoris, and chest pain with exercise), respiratory (chronic cough and trouble getting breath), memory (one *symptom item of* problems with learning or memory), pain (migraine, pain in heart chest, severe headache, and prolonged pain in arms, legs, or back), gastrointestinal (nausea or upset stomach), and fatigue (faintness and feeling weak). In the second step, the 37 *symptom items* were additionally arranged into two *global domains*, including psychological (*symptom items* corresponding to the two *sub-domains* of anxiety and depression) and somatic (*symptom items* corresponding to the eight remaining *sub-domains*). After organizing the symptoms into domains, symptom summaries were generated by counting the total number of symptoms present overall and within each domain, excluding the two sub-domains with only one *symptom item*. This process resulted in the creation of 11 additional cross-sectional measures, leading to a total of 48 cross-sectional symptom measures available at three time-points.

(B) Next, longitudinal symptom patterns hypothesized to affect future HRQoL were engineered from the 48 aforementioned cross-sectional measures. This process, known as feature engineering in machine learning, is a popular technique to utilize domain knowledge for developing new, meaningful features/measures from raw measures. As part of this process, we constructed 10 clinically meaningful longitudinal symptom patterns, labeled as P1 to P10, as follows:

P1. Increase during the first two surveys, regardless of the third
P2. Increase during the last two surveys, regardless of the first
P3. Decrease during the first two surveys, regardless of the third
P4. Decrease during the last two surveys, regardless of the first
P5. Consistent symptom presence at all three surveys
P6. Consistent symptom presence at the first two surveys but not the third
P7. Consistent symptom presence at the last two surveys but not the first
P8. Consistent symptom absence at all three surveys
P9. Consistent symptom absence at the first two surveys but not the third
P10. Consistent symptom absence at the last two surveys but not the first

For the 11 symptom summaries, it is important to note that the consistency patterns were determined based on the presence of any symptoms (i.e., count > 0), whereas the increase/decrease patterns were determined based on the counts of present symptom items. Thus, unlike other patterns, the increase/decrease patterns corresponding to symptom summaries are continuous, ranging from 0 to the maximum number of symptoms in the domain.

Modelling of HRQoL Outcomes Utilizing Clinical Measures and Expert-generated Symptom Patterns

Two distinct modelling attempts were conducted, resulting in the development of two models for each HRQoL outcome. The first model, called the clinical model, utilized clinical data (no symptom data) as potential risk factors. The second model, called the symptom model, integrated symptom patterns alongside the clinical variables.

To explore and identify combinations of risk factors associated with the outcome and assess the strengths of these associations, we utilized Bayesian-Information-Criterion Elastic Net (BIEN).

BIEN is a statistical technique that performs hypothesis testing and estimation simultaneously, allowing us to start without an a priori hypothesis regarding the set of risk factors associated with outcome. Application of BIEN is particularly relevant for this work, given the challenges of formulating clear hypotheses due to the diverse and unique experience risk factors among the heterogeneous population of cancer survivors, the continuously evolving treatments and, thus, their impact on survivors, and the increasing number of newly collected symptoms with limited prior knowledge about them. By employing BIEN, we can effectively account for the joint effects of a wide range of potential risk factors, enabling the identification of risk factors that may be associated exclusively when considered alongside other risk factors. Detailed information on BIEN can be found in the previously published work [Reference to BIEN paper will be provided upon acceptance], and the source code is available online [Link will be provided upon acceptance].

BIEN consists of five key components, which are briefly described as follows:

1) Risk factor set generation: Utilizing Elastic Net, a penalized regression approach, to control the magnitude of the coefficients, this component generates a candidate risk factor set for a given penalty level and form in the optimization function.
2) Model estimation: Employing a maximum likelihood approach, this component estimates the strengths of associations corresponding to the candidate risk factor set generated in 1).
3) Model scoring: This component involves calculating a ranking score for the candidate model generated in 2) using Bayesian Information Criterion. The score takes into account both the goodness of fit and the complexity of the model.
4) Model search: This component conducts a truncated grid search to efficiently suggest a limited set of penalty levels and forms to be investigated by component 1.
5) Model pruning: The final selected model from the preceding components is then pruned by a backward elimination approach, iteratively removing non-significant risk factors.

Evaluation

Finally, to evaluate the predictive performance of the models, we utilized the receiver operating curve (ROC) and calculated the Area Under the Curve (AUC) based on a cutoff of one standard deviation below the normative population mean, which is deemed to be clinically meaningful. The 10-fold cross-validated AUCs were also provided as unbiased measure of performance where sample participants were divide into 10 folds, and modelling was repeated 10 times, with each iteration using 9 folds for model training and the remaining fold for model evaluation.

3.2.4. Analytic Software

All analyses were performed using R version 4.21 (R Project for Statistical Computing). The R function BIEN was used to for modelling [Reference available upon publication of the BIEN paper].

3.3. Results

The 576 5-year survivors in this study were, on average, 9.4 (IQR 4.5–14.1) years old at the time of cancer diagnosis, 27.1 (IQR 23.0-30.4) years old at baseline survey (time-point 1), and 39.7 (IQR 34.8-44.6) years old at time-point 3. Among these survivors, 52% were women, 90% were White race, and 34% had a college degree or higher (Table 3.1). The most common type of cancer was leukemia (41%), followed by Hodgkin lymphoma (20%). In terms of treatment exposure, 89% received chemotherapy, 72% received radiation, and 61% underwent surgery, including 31 survivors who underwent amputation.

Table 3.2 provides detailed information of survivors' symptoms at each time-points. Throughout the study follow-up, a vast majority of survivors (91%) experienced symptoms, with 60% reporting at least one psychological symptom and 87% reporting at least one somatic symptom. The most prevalent symptom sub-domains were sensory, pain, and anxiety, with a prevalence between 50 to 60%. Depression and memory problems followed, with a prevalence between 40 to 50%. The remaining domains of cardiac, fatigue, gastrointestinal, motor, and respiratory problems had a prevalence between 20 to 40%. A domain was deemed prevalent if any symptom within that domain was present at any time during the study follow-up.

Table 3.3 shows detailed information of survivors' symptom patterns. Regarding symptom patterns, 82% of survivors experienced an increase in any of their symptoms during the study follow-up (with 58% in the first two and 66% in the last two time-points). Also, 78% of survivors experienced a decrease in symptoms during the study follow-up (with 59% in the first two and 53% in the last two time-points). Moreover, 57% of survivors experienced a consistent presence of symptoms during two or three subsequent time-points of the study follow-up (with 24% during the first two, 39% during the last two, and 32% during all three

subsequent time-points). Finally, all survivors reported a consistent absence of symptom during two or three subsequent time-points of the study follow-up (with 62% during the first two, 54% during the last two, and 100% during all three subsequent time-points).

Consistently across all ten HRQoL scores, the analysis repeatedly eliminated non-symptom measures when transitioning from clinical models to symptom models, with the exception of age at the 3rd symptom survey and a diagnosis of osteosarcoma, which retained their presence with coefficients similar to those in clinical model. Notably, two frequently selected demographic variables in the clinical models, namely having a college degree or higher (selected seven times) and being a female (selected four times), were completely disregarded in the presence of symptoms. Moreover, specific cancer treatments, which were selected four times in total by clinical models, also disappeared in the presence of symptoms (Figure 3.1: A and B). The transition from clinical to symptom models significantly improved the prediction performance for all ten HRQoL scores, with AUC values ranging from 0.74 to 0.85 in the presence of symptoms compared to 0.56 to 0.66 in their absence (supplementary table 3.1).

Incorporating symptom patterns for MCS rendered all clinical variables non-significant as risk factors, identifying instead eight symptom risk factors, seven of which highlighted the contribution of specific sub-domains of anxiety, depression, fatigue, along with one reflecting overall somatic symptoms (Table 3.4.B). Among these symptom risk factors, the four consistent absence patterns were associated with better MCS, the three consistent presence patterns were associated with poorer MCS, and the one increase pattern was associated with poorer MCS. The selection of three of the eight symptom risk factors was supported by half or more of the cross-validation iterations. The inclusion of symptom patterns improved the

prediction AUC from 0.57 (cross-validation range 0.44-0.73) to 0.81 (cross-validation range 0.63-0.91), as depicted in Figure 3.2 and Supplementary Table 3.1.

In modelling PCS in the absence of symptom data, age at the 3rd time-point and receiving abdomen radiation were identified to be associated with poorer PCS, and having a college degree or higher was identified to be associated with better PCS (Table 3.4.C). However, when considering symptom data, age at the 3rd time-point remained as the only clinical risk factor associated with poorer PCS (Table 3.4.D) along with five symptom patterns, four of which highlighted specific sub-domains of motor, cardiac, pain and fatigue, along with one reflecting overall somatic symptoms. Among these symptom patterns, the four consistent absence were associated with better PCS and the one consistent presence was associated with poorer PCS. The selection of five out of the six risk factors was supported by more than half of the cross-validation iterations. By including symptom patterns, the prediction AUC improved from 0.64 (cross-validation range 0.55-0.76) to 0.83 (cross-validation range 0.70-0.95), as illustrated in Figure 3.2 and Supplementary Table 3.1.

Comparing contribution of symptom sub-domains to MCS and PCS, fatigue symptoms contributed to both MCS and PCS, symptoms of depression and anxiety contributed to MCS, symptoms of motor, cardiac and pain contributed to PCS. Finally, the sub-domains of sensory, respiratory and gastrointestinal, although relevant for certain HRQoL scores, did not appear to contribute to the modelling of either MCS or PCS. As a more comprehensive investigation of symptoms contributing to HRQoL, comparing *symptom items* for the 10 HRQoL outcomes revealed the followings: out of the total 77 risk factors selected, six were clinical risk factors while the remaining 71 were symptom patterns. Among the selected patterns, consistent absence at three time-points had the most major contribution, constituting 40 of the 71 selected

patterns, and consistently associated with better HRQoL. The next frequently selected pattern was consistent presence at all three time-points, constituting 11 of the 71 selected patterns, and consistently associated with poorer HRQoL. The remaining 20 selected patterns spread over five types of patterns as follows: consistent presence over two time-points, including eight consistent presence patterns over the last two time-points and three consistent presence patterns over the first two time-points, increase patterns including six increase patterns during the last two time-points and two increase patterns during the first two time-points, and finally, one consistent absence at the first two time-points. The three remaining types of patterns, including consistent absence over the last two time-points and the two decrease patterns, were never selected to be associated with any HRQoL scores.

3.4. Discussion

Utilizing a diverse set of clinical and longitudinal patient-reported symptom patterns over 3 time points, we modelled the 10 HRQoL scores in survivors of childhood cancer. With the exception of age and osteosarcoma, HRQoL models relied on symptom patterns as associative factors of HRQoL. The exclusion of the other clinical variables suggests that symptoms may mediate their relationship with HRQoL. Incorporating symptom patterns significantly improved the prediction performance for future HRQoL, indicating the valuable insights they could offer. These findings could have important implications guiding symptom surveillance during follow-ups and symptom management interventions to alleviate specific symptoms.

While there is an extensive body of literature on HRQoL in childhood cancer survivors, this study has the potential to uncover valuable additional insight on avenues to improve long-term well-being of childhood cancer survivors through emphasizing several key issues. First, by focusing on symptoms as the primary source of risk factors, this study aims to identify actionable intervention targets that can be addressed to enhance well-being. Second, by utilizing various longitudinal symptom patterns, it acknowledges the dynamic nature of symptoms over time and their varying impact. Third, the data-driven selection of risk factors in this study allows for timely and novel extraction of risk factors from a large pool, particularly useful in rapidly-evolving areas with limited literature to guide the selection. Fourth, its particular emphasis on patient-reported measures could contribute to patient-centered care by underscoring the patient`s voice. Finally, the statistical tool employed considers the interrelationship among candidate risk factors, facilitating the identification of risk factors that work well in combination.

The observed appreciable contribution of symptom data in modelling HRQoL, along with the reduced impact of clinical variables, aligns with previous research suggesting the direct impact of symptoms on HRQoL beyond clinical variables [7, 85]. This finding supports the importance of collecting patient-generated symptoms from childhood cancer survivors and motivates the implementation of surveillance programs for lifelong symptom monitoring, leveraging innovative methods like wearable or mobile devices.

The symptoms identified in this study as being associated with HRQoL can inform the survivors about the appropriate time to refer to clinics and facilitate the communication between clinicians and survivors during the clinical visit, aiding clinicians in making appropriate referral decisions. Furthermore, these symptoms can inform the design of future clinical trials.

The selection of diverse patterns in the analysis highlights the importance of incorporating a broad array of pattern types for symptoms. First, the selection of consistent absence/presence over all three time-points, notably more often than all other patterns, suggests them as a valuable source of information. Not surprisingly, the robustness of these patterns is supported by their more consistent selection across cross-validation iterations compared to other patterns. Second, the higher selection of patterns of consistent presence or increase during the last two time-points, compared to the first two time-points, suggests a worsening state or consistent presence of symptoms closer to the outcome assessment could be more impactful, while similar pattern occurring earlier in the timeframe may allow individuals more time to adapt.

Several limitations should be acknowledged regarding this study. Firstly, the majority of the study survivors are white (90%) and were exclusively recruited from a single institution. Consequently, the generalizability of the findings to the other childhood cancer survivors

treated at different institutes may be limited. It would be valuable to obtain large-scale data from other institutes to assess the applicability of the model in a more diverse survivor population. Another potential caveat in this study is that the time-points of symptom assessment were dictated by the specific time-points of surveys that included symptom questions. Incorporating continuous symptom monitoring through activity trackers or smartphones could offer opportunities for further advancement in this area of research. Furthermore, two of the eight HRQoL scores involved in the calculation of MCS and PCS, namely bodily pain and mental health scores, encompass both symptom status and functional status, creating a potential overlap between these scores and symptom data. Nonetheless, the good performance for the remaining six HRQoL scores suggests that comparable performance for MCS and PCS can likely be achieved, despite the aforementioned overlap. Moreover, inherent to cohort studies featuring long-term follow-ups, our findings may be vulnerable to selection bias arising from non-participation and non-response, raising concerns about the validity of our results. Addressing non-participation among eligible participants poses challenges due to the lack of comprehensive data on non-participants' outcomes and exposures. While prior analysis found no substantial differences in demographic and cancer-related characteristics between participants and non-participants of the two cohort studies, alleviating serious concerns regarding selective non-participation in the cohorts, it's plausible that other variables still differ between the two groups. Non-response among initially consenting participants might is also of concern due to potential differences in symptoms and HRQoL between responders and non-responders: survivors might opt out due to deteriorating health or due to doing well and not wanting to be reminded of their cancer experience, leading to an overly optimistic or pessimistic portrayal of survivor experiences compared to the typical

survivors, respectively. Given these potential biases, caution is warranted in interpreting our results.

This study is the first to investigate the associations between hundreds of patterns of patient-reported symptoms and future HRQoL in childhood cancer survivors over an extended period. Our findings align with existing literature, highlighting the importance of continuous symptom monitoring in this population. We identified that symptoms related to anxiety, depression, and fatigue were associated with MCS, and symptoms related to motor, cardiac, pain and fatigue were associated with PCS. The identified associative symptoms hold promise to improve survivors` lives by facilitating timely medical appointments and counselling services to those experiencing certain risk factors, reducing the burden of late or unnecessary services. Further, during the service provision, such information can trigger helpful communication between clinicians and survivors, aiding shared decision-making. Additionally, it has implications for future evidence-based clinical practice guidelines. As the number of long-term cancer survivors and routinely collected symptoms continues to rise, leveraging available data through research like this becomes crucial for better HRQoL outcomes. Our analysis framework can serve as a valuable pipeline across diverse clinical settings.

Table 3.1: Demographic, cancer diagnosis, and treatment characteristics of study participants

Variable	Number (%) or Mean (Standard Deviation)
Demographic data	
1 Age at Baseline Survey [Year]	27.1 (5.02)
2 Time between Baseline and Outcome Survey [Year]	15.4 (6.15)
3 Age at 3rd Symptom Survey [Year]	39.7 (7.39)
4 Time between 3rd Symptom and Outcome Survey [Year]	2.8 (1.74)
5 Age at Diagnosis of Cancer [Year]	9.4 (5.42)
6 Sex (Female)	298 (51.7%)
7 Race (White)	517 (89.8%)
8 Educational Attainment at Baseline Survey (College graduate or higher)	193 (33.5%)
Cancer Diagnosis	
9 Leukemia	234 (40.6%)
10 Hodgkin lymphoma	115 (20.0%)
11 Non-Hodgkin lymphoma	55 (9.5%)
12 Osteosarcoma	44 (7.6%)
13 Wilms tumor	36 (6.2%)
14 Neuroblastoma	24 (4.2%)
15 Central nervous system tumors	29 (5.0%)
16 Other malignancy	39 (6.8%)
Cancer Treatment	
Chemotherapy Exposure	
17 Methotrexate	307 (53.3%)
18 Intrathecal Methotrexate	255 (44.3%)
19 High dose Methotrexate	132 (22.9%)
20 Cytarabine	177 (30.7%)
21 Intrathecal Cytarabine	105 (18.2%)
22 High dose Cytarabine	20 (3.5%)
23 Bleomycin	35 (6.1%)
24 Alkylating agent	373 (64.8%)
25 Anthracycline	342 (59.4%)
26 Corticosteroid	319 (55.4%)
27 Plant alkaloid	448 (77.8%)
28 Platinum	37 (6.4%)
Radiation Exposure	
29 Brain radiation	234 (40.6%)
30 Neck radiation	187 (32.5%)
31 Chest radiation	192 (33.3%)
32 Abdomen radiation	161 (28.0%)
33 Pelvis radiation	140 (24.3%)
Surgery	
34 Amputation	31 (5.4%)
35 Other surgery	320 (55.6%)

Table 3.2: Symptom characteristics of study participants at three survey time-points.

Variable (Category: subcategory)	Number (%) or Mean (Standard Deviation) at T1 (~1996)	T2 (~2008)	T3 (~2013)
Psychological Symptoms			
1 *Depression:* Thoughts of ending life	9 (1.6%)	8 (1.4%)	5 (0.9%)
2 *Depression:* Feeling lonely	81 (14.1%)	78 (13.5%)	99 (17.2%)
3 *Depression:* Feeling blue	85 (14.8%)	87 (15.1%)	110 (19.1%)
4 *Depression:* Feeling no interest in things	55 (9.5%)	82 (14.2%)	88 (15.3%)
5 *Depression:* Feeling hopeless about the future	50 (8.7%)	63 (10.9%)	71 (12.3%)
6 *Depression:* Feelings of worthlessness	37 (6.4%)	43 (7.5%)	59 (10.2%)
7 *Depression:* Summary of Items*	0.55 (1.24)	0.63 (1.35)	0.75 (1.41)
8 *Anxiety:* Nervousness or shaking inside	54 (9.4%)	64 (11.1%)	86 (14.9%)
9 *Anxiety:* Suddenly scared for no reason	29 (5.0%)	30 (5.2%)	37 (6.4%)
10 *Anxiety:* Feeling fearful	47 (8.2%)	50 (8.7%)	55 (9.5%)
11 *Anxiety:* Feeling tense or keyed up	112 (19.4%)	114 (19.8%)	131 (22.7%)
12 *Anxiety:* Spells of terror or panic	30 (5.2%)	35 (6.1%)	43 (7.5%)
13 *Anxiety:* So restless cannot sit still	63 (10.9%)	75 (13.0%)	65 (11.3%)
14 *Anxiety:* Summary of Items*	0.58 (1.20)	0.64 (1.28)	0.72 (1.36)
Somatic Symptoms			
15 *Sensory:* Decreased sense of touch	27 (4.7%)	44 (7.6%)	54 (9.4%)
16 *Sensory:* Tinnitus/ringing in ear	39 (6.8%)	57 (9.9%)	80 (13.9%)
17 *Sensory:* Dizziness	21 (3.6%)	23 (4.0%)	35 (6.1%)
18 *Sensory:* Double vision	13 (2.3%)	7 (1.2%)	12 (2.1%)
19 *Sensory:* Other trouble seeing	22 (3.8%)	24 (4.2%)	47 (8.2%)
20 *Sensory:* Very dry eyes	43 (7.5%)	50 (8.7%)	52 (9.0%)
21 *Sensory:* Abnormal Sense of taste	11 (1.9%)	7 (1.2%)	8 (1.4%)
22 *Sensory:* Numbness	48 (8.3%)	78 (13.5%)	108 (18.8%)
23 *Sensory:* Summary of Items*	0.39 (0.80)	0.50 (0.87)	0.69 (1.03)
24 Motor: Problem with balance	35 (6.1%)	51 (8.9%)	75 (13.0%)
25 Motor: Tremors/movement problems	19 (3.3%)	13 (2.3%)	22 (3.8%)
26 Motor: Weakness/inability to move arm	27 (4.7%)	24 (4.2%)	33 (5.7%)
27 Motor: Weakness/inability to move leg	26 (4.5%)	18 (3.1%)	30 (5.2%)
28 Motor: Summary of Items*	0.19 (0.61)	0.18 (0.54)	0.28 (0.70)
29 Cardiac: Arrhythmia	28 (4.9%)	40 (6.9%)	55 (9.5%)
30 Cardiac: Angina pectoris	2 (0.3%)	3 (0.5%)	7 (1.2%)
31 Cardiac: Chest pain with exercise	72 (12.5%)	65 (11.3%)	96 (16.7%)
32 Cardiac: Summary of Items*	0.18 (0.43)	0.19 (0.48)	0.27 (0.57)
33 Respiratory: Chronic cough	34 (5.9%)	31 (5.4%)	39 (6.8%)
34 Respiratory: Trouble getting breath	33 (5.7%)	31 (5.4%)	57 (9.9%)
35 Respiratory: Summary of Items*	0.12 (0.37)	0.11 (0.37)	0.17 (0.45)
36 Memory: Problems with learning or memory	76 (13.2%)	129 (22.4%)	173 (30.0%)
37 Pain: Migraine	97 (16.8%)	91 (15.8%)	97 (16.8%)
38 Pain: Pain in heart chest	20 (3.5%)	36 (6.2%)	44 (7.6%)
39 Pain: Severe headache	110 (19.1%)	83 (14.4%)	65 (11.3%)
40 Pain: Prolonged pain in arms, legs, or back	76 (13.2%)	97 (16.8%)	117 (20.3%)
41 Pain: Summary of Items*	0.53 (0.83)	0.53 (0.86)	0.56 (0.86)
42 Gastrointestinal: Nausea or upset stomach	74 (12.8%)	68 (11.8%)	75 (13.0%)
43 Fatigue: Faintness	19 (3.3%)	27 (4.7%)	49 (8.5%)
44 Fatigue: Feeling weak	50 (8.7%)	72 (12.5%)	92 (16.0%)
45 Fatigue: Summary of Items*	0.12 (0.37)	0.17 (0.44)	0.24 (0.52)
Psychological and/or Somatic Summaries			
46 Psychological: Summary of Items*	1.13 (2.22)	1.27 (2.38)	1.47 (2.48)
47 Somatic: Summary of Items*	1.77 (2.53)	2.03 (2.70)	2.64 (3.23)
48 Psychological/Somatic: Summary of Items*	2.91 (4.02)	3.30 (4.33)	4.12 (4.96)

Note: * Continuous measures counting positive symptom items in the corresponding domain

Table 3.3: Longitudinal symptom pattern characteristics of study participants over three survey time-points.

Variable (Category: **subcategory)**	**Number (%) or Mean (Standard Deviation) for Patterns† of**									
	P1: -,+,+/-	**P2: +/-,-,+**	**P3: +,-,+/-**	**P4: +/-,+,-**	**P5: +,+,+**	**P6: +,+,-**	**P7: -,+,+**	**P8: -,-,-**	**P9**: -,-,+	**P10**: +,-,-
Psychological Symptoms										
1 Depression: Thoughts of ending life	5 (0.9%)	5 (0.9%)	6 (1.0%)	8 (1.4%)	0 (0.0%)	3 (0.5%)	0 (0.0%)	557 (96.7%)	5 (0.9%)	6 (1.0%)
2 Depression: Feeling lonely	51 (8.9%)	54 (9.4%)	54 (9.4%)	33 (5.7%)	18 (3.1%)	9 (1.6%)	27 (4.7%)	403 (70.0%)	41 (7.1%)	41 (7.1%)
3 Depression: Feeling blue	57 (9.9%)	68 (11.8%)	55 (9.5%)	45 (7.8%)	18 (3.1%)	12 (2.1%)	24 (4.2%)	381 (66.1%)	53 (9.2%)	40 (6.9%)
4 Depression: Feeling no interest in things	57 (9.9%)	47 (8.2%)	30 (5.2%)	41 (7.1%)	15 (2.6%)	10 (1.7%)	26 (4.5%)	428 (74.3%)	36 (6.2%)	19 (3.3%)
5 Depression: Feeling hopeless about the future	45 (7.8%)	43 (7.5%)	32 (5.6%)	35 (6.1%)	11 (1.9%)	7 (1.2%)	17 (3.0%)	443 (76.9%)	38 (6.6%)	27 (4.7%)
6 Depression: Feelings of worthlessness	32 (5.6%)	38 (6.6%)	26 (4.5%)	22 (3.8%)	9 (1.6%)	2 (0.3%)	12 (2.1%)	475 (82.5%)	32 (5.6%)	20 (3.5%)
7 Depression: Summary of Items*	0.40 (1.03)	0.41 (1.01)	0.32 (0.93)	0.28 (0.82)	42 (7.3%)	18 (3.1%)	45 (7.8%)	316 (54.9%)	57 (9.9%)	43 (7.5%)
8 Anxiety: Nervousness or shaking inside	41 (7.1%)	53 (9.2%)	31 (5.4%)	31 (5.4%)	14 (2.4%)	9 (1.6%)	19 (3.3%)	433 (75.2%)	48 (8.3%)	26 (4.5%)
9 Anxiety: Suddenly scared for no reason	17 (3.0%)	25 (4.3%)	16 (2.8%)	18 (3.1%)	6 (1.0%)	7 (1.2%)	6 (1.0%)	508 (88.2%)	22 (3.8%)	13 (2.3%)
10 Anxiety: Feeling fearful	36 (6.2%)	29 (5.0%)	33 (5.7%)	24 (4.2%)	12 (2.1%)	2 (0.3%)	14 (2.4%)	467 (81.1%)	26 (4.5%)	30 (5.2%)
11 Anxiety: Feeling tense or keyed up	66 (11.5%)	73 (12.7%)	64 (11.1%)	56 (9.7%)	32 (5.6%)	16 (2.8%)	26 (4.5%)	342 (59.4%)	56 (9.7%)	47 (8.2%)
12 Anxiety: Spells of terror or panic	24 (4.2%)	26 (4.5%)	19 (3.3%)	18 (3.1%)	9 (1.6%)	2 (0.3%)	8 (1.4%)	498 (86.5%)	24 (4.2%)	17 (3.0%)
13 Anxiety: So restless cannot sit still	49 (8.5%)	37 (6.4%)	37 (6.4%)	47 (8.2%)	14 (2.4%)	12 (2.1%)	14 (2.4%)	433 (75.2%)	31 (5.4%)	31 (5.4%)
14 Anxiety: Summary of Items*	0.38 (0.95)	0.39 (0.92)	0.32 (0.79)	0.30 (0.80)	56 (9.7%)	28 (4.9%)	40 (6.9%)	273 (47.4%)	63 (10.9%)	57 (9.9%)
Somatic Symptoms										
15 Sensory: Decreased sense of touch	33 (5.7%)	32 (5.6%)	16 (2.8%)	22 (3.8%)	6 (1.0%)	5 (0.9%)	16 (2.8%)	487 (84.5%)	29 (5.0%)	13 (2.3%)
16 Sensory: Tinnitus/ringing in ear	36 (6.2%)	37 (6.4%)	18 (3.1%)	14 (2.4%)	17 (3.0%)	4 (0.7%)	26 (4.5%)	468 (81.2%)	33 (5.7%)	14 (2.4%)
17 Sensory: Dizziness	14 (2.4%)	26 (4.5%)	12 (2.1%)	14 (2.4%)	5 (0.9%)	4 (0.7%)	4 (0.7%)	517 (89.8%)	24 (4.2%)	10 (1.7%)
18 Sensory: Double vision	3 (0.5%)	9 (1.6%)	9 (1.6%)	4 (0.7%)	3 (0.5%)	1 (0.2%)	0 (0.0%)	553 (96.0%)	7 (1.2%)	7 (1.2%)
19 Sensory: Other trouble seeing	15 (2.6%)	35 (6.1%)	13 (2.3%)	12 (2.1%)	5 (0.9%)	4 (0.7%)	7 (1.2%)	507 (88.0%)	32 (5.6%)	10 (1.7%)
20 Sensory: Very dry eyes	34 (5.9%)	30 (5.2%)	27 (4.7%)	28 (4.9%)	8 (1.4%)	8 (1.4%)	14 (2.4%)	477 (82.8%)	22 (3.8%)	19 (3.3%)
21 Sensory: Abnormal Sense of taste	4 (0.7%)	6 (1.0%)	8 (1.4%)	5 (0.9%)	0 (0.0%)	3 (0.5%)	2 (0.3%)	555 (96.4%)	6 (1.0%)	8 (1.4%)
22 Sensory: Numbness	59 (10.2%)	69 (12.0%)	29 (5.0%)	39 (6.8%)	16 (2.8%)	3 (0.5%)	23 (4.0%)	408 (70.8%)	61 (10.6%)	21 (3.6%)
23 Sensory: Summary of Items*	0.29 (0.64)	0.38 (0.73)	0.18 (0.47)	0.20 (0.50)	67 (11.6%)	24 (4.2%)	57 (9.9%)	240 (41.7%)	86 (14.9%)	31 (5.4%)
24 Motor: Problem with balance	32 (5.6%)	41 (7.1%)	16 (2.8%)	17 (3.0%)	15 (2.6%)	4 (0.7%)	19 (3.3%)	470 (81.6%)	39 (6.8%)	14 (2.4%)
25 Motor: Tremors/movement problems	11 (1.9%)	19 (3.3%)	17 (3.0%)	10 (1.7%)	1 (0.2%)	1 (0.2%)	2 (0.3%)	529 (91.8%)	17 (3.0%)	15 (2.6%)
26 Motor: Weakness/inability to move arm	10 (1.7%)	21 (3.6%)	13 (2.3%)	12 (2.1%)	8 (1.4%)	6 (1.0%)	4 (0.7%)	522 (90.6%)	17 (3.0%)	9 (1.6%)
27 Motor: Weakness/inability to move leg	10 (1.7%)	23 (4.0%)	18 (3.1%)	11 (1.9%)	5 (0.9%)	3 (0.5%)	2 (0.3%)	522 (90.6%)	18 (3.1%)	13 (2.3%)
28 Motor: Summary of Items*	0.10 (0.36)	0.17 (0.56)	0.10 (0.40)	0.07 (0.32)	24 (4.2%)	10 (1.7%)	25 (4.3%)	422 (73.3%)	46 (8.0%)	24 (4.2%)
29 Cardiac: Arrhythmia	25 (4.3%)	26 (4.5%)	13 (2.3%)	11 (1.9%)	12 (2.1%)	3 (0.5%)	17 (3.0%)	500 (86.8%)	23 (4.0%)	10 (1.7%)
30 Cardiac: Angina pectoris	3 (0.5%)	6 (1.0%)	2 (0.3%)	2 (0.3%)	0 (0.0%)	0 (0.0%)	1 (0.2%)	565 (98.1%)	6 (1.0%)	2 (0.3%)
31 Cardiac: Chest pain with exercise	38 (6.6%)	53 (9.2%)	45 (7.8%)	22 (3.8%)	23 (4.0%)	4 (0.7%)	20 (3.5%)	430 (74.7%)	36 (6.2%)	28 (4.9%)
32 Cardiac: Summary of Items*	0.11 (0.35)	0.15 (0.40)	0.10 (0.31)	0.06 (0.25)	33 (5.7%)	6 (1.0%)	27 (4.7%)	391 (67.9%)	46 (8.0%)	34 (5.9%)
33 Respiratory: Chronic cough	17 (3.0%)	22 (3.8%)	20 (3.5%)	14 (2.4%)	6 (1.0%)	8 (1.4%)	11 (1.9%)	506 (87.8%)	19 (3.3%)	17 (3.0%)
34 Respiratory: Trouble getting breath	24 (4.2%)	38 (6.6%)	26 (4.5%)	12 (2.1%)	5 (0.9%)	2 (0.3%)	14 (2.4%)	485 (84.2%)	34 (5.9%)	22 (3.8%)
35 Respiratory: Summary of Items*	0.07 (0.29)	0.10 (0.35)	0.07 (0.28)	0.05 (0.23)	15 (2.6%)	7 (1.2%)	17 (3.0%)	452 (78.5%)	38 (6.6%)	26 (4.5%)
36 Memory: Problems with learning or memory	83 (14.4%)	77 (13.4%)	30 (5.2%)	33 (5.7%)	39 (6.8%)	7 (1.2%)	57 (9.9%)	346 (60.1%)	71 (12.3%)	24 (4.2%)
37 Pain: Migraine	39 (6.8%)	35 (6.1%)	45 (7.8%)	29 (5.0%)	39 (6.8%)	13 (2.3%)	23 (4.0%)	415 (72.0%)	25 (4.3%)	35 (6.1%)
38 Pain: Pain in heart chest	30 (5.2%)	33 (5.7%)	14 (2.4%)	25 (4.3%)	3 (0.5%)	3 (0.5%)	8 (1.4%)	498 (86.5%)	28 (4.9%)	9 (1.6%)

Variable (Category: **subcategory)**	**Number (%) or Mean (Standard Deviation) for Patterns† of**									
	P1: -,+,+/-	**P2: +/-,-,+**	**P3: +,-,+/-**	**P4: +/-,+,-**	**P5: +,+,+**	**P6: +,+,-**	**P7: -,+,+**	**P8: -,-,-**	**P9**: -,-,+	**P10**: +,-,-
39 Pain: Severe headache	40 (6.9%)	29 (5.0%)	67 (11.6%)	47 (8.2%)	21 (3.6%)	22 (3.8%)	15 (2.6%)	412 (71.5%)	14 (2.4%)	52 (9.0%)
40 Pain: Prolonged pain in arms, legs, or back	54 (9.4%)	59 (10.2%)	33 (5.7%)	39 (6.8%)	30 (5.2%)	13 (2.3%)	28 (4.9%)	401 (69.6%)	45 (7.8%)	19 (3.3%)
41 Pain: Summary of Items*	0.24 (0.59)	0.23 (0.50)	0.23 (0.51)	0.20 (0.52)	96 (16.7%)	34 (5.9%)	42 (7.3%)	255 (44.3%)	47 (8.2%)	42 (7.3%)
42 Gastrointestinal: Nausea or upset stomach	47 (8.2%)	51 (8.9%)	53 (9.2%)	44 (7.6%)	13 (2.3%)	8 (1.4%)	11 (1.9%)	421 (73.1%)	34 (5.9%)	36 (6.2%)
43 Fatigue: Faintness	22 (3.8%)	37 (6.4%)	14 (2.4%)	15 (2.6%)	1 (0.2%)	4 (0.7%)	11 (1.9%)	500 (86.8%)	35 (6.1%)	12 (2.1%)
44 Fatigue: Feeling weak	47 (8.2%)	50 (8.7%)	25 (4.3%)	30 (5.2%)	20 (3.5%)	5 (0.9%)	22 (3.8%)	436 (75.7%)	43 (7.5%)	18 (3.1%)
45 Fatigue: Summary of Items*	0.11 (0.35)	0.14 (0.41)	0.06 (0.25)	0.07 (0.28)	23 (4.0%)	7 (1.2%)	30 (5.2%)	409 (71.0%)	51 (8.9%)	19 (3.3%)
Psychological and/or Somatic Summaries										
46 Psychological: Summary of Items*	0.73 (1.74)	0.74 (1.66)	0.60 (1.49)	0.53 (1.40)	88 (15.3%)	30 (5.2%)	59 (10.2%)	233 (40.5%)	56 (9.7%)	48 (8.3%)
47 Somatic: Summary of Items*	0.88 (1.68)	1.12 (1.90)	0.63 (1.25)	0.51 (1.16)	238 (41.3%)	27 (4.7%)	72 (12.5%)	77 (13.4%)	50 (8.7%)	38 (6.6%)
48 Psychological/Somatic: Summary of Items*	1.46 (2.75)	1.67 (2.91)	1.07 (2.24)	0.85 (1.94)	279 (48.4%)	36 (6.2%)	62 (10.8%)	51 (8.9%)	46 (8.0%)	33 (5.7%)

Note: * For summary measures, P1-P4 (i.e., increase/decrease) patterns are continuous determined based on the counts of present symptom items, while P5-P10 (i.e., consistency patterns) are binary and determined based on the presence of any symptoms (i.e., count > 0), † Patterns, denoted as P1-P10, are characterized by three signs indicators representing symptom status at T1, T2, and T3, where + indicates symptom presence, - indicates symptom absence, and +/- indicates that symptom may be present or absent. P1-P10 are described as follows: P1 (Increase from T1 to T2); P2 (Increase from T2 to T3); P3 (Decrease from T1 to T2); P4 (Decrease from T2 to T3); P5 (Consistent Presence at T1, T2, & T3); P6 (Consistent Presence at T1 & T2 but not T3); P7 (Consistent Presence at T2 & T3 but not T1); P8 (Consistent Absence at T1, T2, & T3); P9 (Consistent Absence at T1 & T2 but not T3); P10 (Consistent Absence at T2 & T3 but not T1).

Table 3.4: Selected models by BIEN for Mental and Physical Component Score outcomes without and with symptom patterns.

Mental Component Score Outcome					
A. Clinical Model (selecting risk factors from the 35 clinical variables)					
Risk Factor Name		**Estimate**	**95% CI***	**P-value**	**Support†**
(Intercept)		50.13	48.81, 51.45	<0.001	
Demographic: Sex (Female)		-3.24	-5.07, -1.4	<0.001	100%
B. Symptom Model (selecting risk factors from the 35 clinical variables + 480 symptom patterns)					
Risk Factor Name ("*Symptom Sub-Domain***:** Symptom Item or **Summary of Items" or "<u>*Symptom Global Domain*</u>: Summary of Items")**	**Pattern Type**	**Estimate**	**95% CI***	**P-value**	**Support†**
(Intercept)		40.65	38.09, 43.22	<0.001	
Depression: Feeling no interest in things	P8: Consistent Absence at T1, T2, & T3	5.00	2.97, 7.03	<0.001	70%
Depression: Feeling hopeless about the future	P2: Increase from T2 to T3	-7.09	-10.06, -4.11	<0.001	10%
Depression: **Summary of Items ***	P5: Consistent Presence at T1, T2, & T3	-5.14	-8.35, -1.94	0.002	40%
Anxiety: Suddenly scared for no reason	P6: Consistent Presence at T1 & T2 but not T3	-14.07	-21.15, -7.00	<0.001	80%
Anxiety: Feeling tense or keyed up	P8: Consistent Absence at T1, T2, & T3	2.85	1.02, 4.68	0.002	50%
Anxiety: So restless cannot sit still	P8: Consistent Absence at T1, T2, & T3	3.00	0.96, 5.04	0.004	40%
Fatigue: Feeling weak	P8: Consistent Absence at T1, T2, & T3	2.84	0.85, 4.84	0.005	40%
<u>Somatic</u>: **Summary of Items ***	P5: Consistent Presence at T1, T2, & T3	-2.27	-3.92, -0.61	0.007	10%
Physical Component Score Outcome					
C. Clinical Model (selecting risk factors from the 35 clinical variables)					
Risk Factor Name		**Estimate**	**95% CI***	**P-value**	**Support†**
(Intercept)		61.09	56.3, 65.89	<0.001	
Demographic: Age at 3rd Symptom Survey [Year]		-0.35	-0.46, -0.23	<0.001	100%
Demographic: Educational Attainment at Baseline Survey (College graduate or higher)		3.59	1.75, 5.43	<0.001	100%
Treatment: Abdomen radiation		-2.78	-4.72, -0.84	0.005	50%
D. Symptom Model (selecting risk factors from the 35 clinical variables + 480 symptom patterns)					
Risk Factor Name ("*Symptom Sub-Domain***:** Symptom Item or **Summary of Items" or "<u>*Symptom Global Domain*</u>: Summary of Items")**	**Pattern Type**	**Estimate**	**95% CI***	**P-value**	**Support†**
(Intercept)		46.10	41.06, 51.14	<0.001	
Demographic: Age at 3rd Time-point [Year]		-0.33	-0.43, -0.23	<0.001	100%
Motor: Weakness/inability to move leg	P8: Consistent Absence at T1, T2, & T3	5.82	3.12, 8.52	<0.001	80%
Cardiac: Chest pain with exercise	P8: Consistent Absence at T1, T2, & T3	5.41	3.54, 7.28	<0.001	100%
Pain: Prolonged pain in arms, legs, or back	P8: Consistent Absence at T1, T2, & T3	2.91	1.11, 4.71	0.002	80%
Fatigue: Feeling weak	P8: Consistent Absence at T1, T2, & T3	6.01	4.07, 7.95	<0.001	100%
***<u>Somatic</u>*: Summary of Items**	P5: Consistent Presence at T1, T2, & T3	-2.37	-4.04, -0.70	0.006	40%

Note: * Confidence Interval, † Percentage of times the variable was selected in the 10 cross-validation iterations

A) Modelling mental outcomes selecting risk factors from the 35 clinical risk factors

Mental Outcomes: Mental component score	Mental health	Emotional role limitation	Social Functioning	Vitality	Risk Factor Name
		-0.25			*Demographic:* Age at 3rd Symptom Survey [Year]
-3.24	-2.90		-3.01	-3.02	*Demographic:* Sex (Female)
		3.26		3.25	*Demographic:* Educational Attainment at Baseline Survey (College graduate or higher)

B) Modelling physical outcomes selecting risk factors from the 35 clinical risk factors

Physical Outcomes: Physical component score	General health	Physical role limitation	Physical functioning	Bodily pain	Risk Factor Name
				-0.19	*Demographic:* Time between Baseline and Outcome Survey [Year]
-0.35	-0.28	-0.32	-0.32		*Demographic:* Age at 3rd Symptom Survey [Year]
3.59	3.31	3.04	3.50	2.91	*Demographic:* Educational Attainment at Baseline Survey (college graduate or higher)
			-4.46		*Diagnosis:* Osteosarcoma
	-2.87				*Treatment:* Alkylating agent
-2.78		-2.67			*Treatment:* Abdomen radiation
	-3.81				*Treatment:* Pelvis radiation

C) Modelling mental outcomes selecting risk factors from the 35 clinical + 480 symptom pattern risk factors

Mental

Mental component score	Mental health	Emotional role limitation	Social functioning	Vitality	Risk Factor Name	Pattern Type
		-0.27			*Demographic:* Age at 3rd Symptom Survey [Year]	
5.00		3.93		4.31	*Depression*: Feeling no interest in things	P8: Consistent Absence at T1, T2, & T3
-7.09					*Depression*: Feeling hopeless about the future	P2: Increase from T2 to T3
		-5.01			*Depression*: Feeling hopeless about the future	P9: Consistent Absence at T1 & T2 but not T3
	4.12	3.45	4.61		*Depression*: Feelings of worthlessness	P8: Consistent Absence at T1, T2, & T3
				-1.83	*Depression*: Summary of Items	P2: Increase from T2 to T3
-5.14	-5.57				*Depression*: Summary of Items	P5: Consistent Presence at T1, T2, & T3
		-6.34			*Anxiety*: Nervousness or shaking inside	P7: Consistent Presence at T2 & T3 but not T1
-14.07	-18.93		-10.31		*Anxiety*: Suddenly scared for no reason	P6: Consistent Presence at T1 & T2 but not T3
			-6.00		*Anxiety*: Feeling tense or keyed up	P5: Consistent Presence at T1, T2, & T3
2.85	3.14		3.28		*Anxiety*: Feeling tense or keyed up	P8: Consistent Absence at T1, T2, & T3
3.00	3.43	3.33			*Anxiety*: So restless cannot sit still	P8: Consistent Absence at T1, T2, & T3
	-9.04				*Sensory*: Decreased sense of touch	P7: Consistent Presence at T2 & T3 but not T1
			-6.11		*Motor*: Tremors/movement problems	P2: Increase from T2 to T3
		-10.97			*Motor*: Weakness/inability to move arm	P1: Increase from T1 to T2
		-11.00			*Motor*: Weakness/inability to move leg	P1: Increase from T1 to T2
			3.24		*Motor*: Summary of Items	P8: Consistent Absence at T1, T2, & T3
		-7.84			*Cardiac*: Arrhythmia	P7: Consistent Presence at T2 & T3 but not T1
			-7.54		*Cardiac*: Chest pain with exercise	P7: Consistent Presence at T2 & T3 but not T1
		-9.05			*Gastrointestinal*: Nausea or upset stomach	P7: Consistent Presence at T2 & T3 but not T1
2.84	2.95		4.70	4.57	*Fatigue*: Feeling weak	P8: Consistent Absence at T1, T2, & T3
		-7.97			*Fatigue*: Summary of Items	P5: Consistent Presence at T1, T2, & T3
-2.27		-2.78		-4.76	*Somatic*: Summary of Items	P5: Consistent Presence at T1, T2, & T3
	-0.43	-0.56			*Psychological/Somatic*: Summary of Items	P2: Increase from T2 to T3

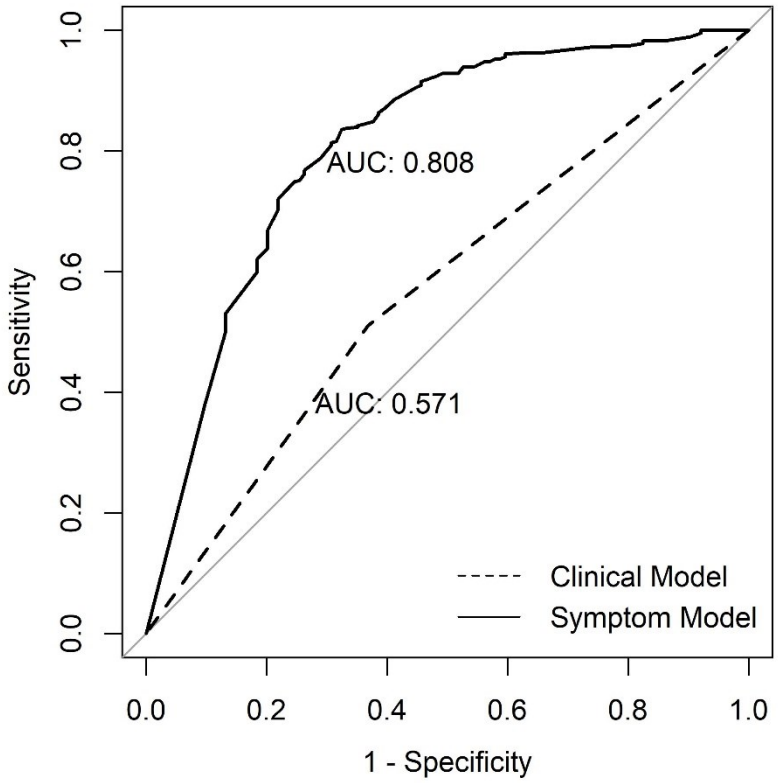
D) Modelling physical outcomes selecting risk factors from the 35 clinical + 480 symptom pattern risk factors

Physical Outcomes: Physical component score	General health	Physical role limitation	Physical functioning	Bodily pain	Risk Factor Name	Pattern Type
-0.33	-0.29	-0.30	-0.25		*Demographic*: Age at 3rd Symptom Survey [Year]	
			-5.49		*Diagnosis*: Osteosarcoma	
		3.67			*Depression*: Feelings of worthlessness	P8: Consistent Absence at T1, T2, & T3
	2.83				*Depression*: Summary of Items	P8: Consistent Absence at T1, T2, & T3
	3.07				*Anxiety*: Nervousness or shaking inside	P8: Consistent Absence at T1, T2, & T3
				2.81	*Anxiety*: Feeling tense or keyed up	P8: Consistent Absence at T1, T2, & T3
			3.46	4.02	*Sensory:* Decreased sense of touch	P8: Consistent Absence at T1, T2, & T3
			-15.59		*Sensory:* Dizziness	P7: Consistent Presence at T2 & T3 but not T1
5.82					*Motor:* Weakness/inability to move leg	P8: Consistent Absence at T1, T2, & T3
		3.54	2.36		*Motor:* Summary of Items	P8: Consistent Absence at T1, T2, & T3
		-7.89	-6.10		*Cardiac:* Chest pain with exercise	P7: Consistent Presence at T2 & T3 but not T1
5.41	5.30	2.65	5.05		*Cardiac:* Chest pain with exercise	P8: Consistent Absence at T1, T2, & T3
	4.29		4.06		*Respiratory:* Chronic cough	P8: Consistent Absence at T1, T2, & T3
2.91		3.34		3.85	*Pain:* Prolonged pain in arms, legs, or back	P8: Consistent Absence at T1, T2, & T3
6.01	6.95	5.99	6.27	3.86	*Fatigue:* Feeling weak	P8: Consistent Absence at T1, T2, & T3
		-6.21			*Fatigue:* Summary of Items	P5: Consistent Presence at T1, T2, & T3
-2.37	-2.77			-3.76	*Somatic:* Summary of Items	P5: Consistent Presence at T1, T2, & T3
				-0.51	*Psychological/Somatic:* Summary of Items	P2: Increase from T2 to T3

Figure 3.1: Selected risk factors and estimated coefficients for HRQoL Models.

Selected risk factors and estimated coefficients for 576 childhood cancer survivors in A) clinical models for mental outcomes, B) clinical models for physical outcomes, C) symptom models for mental outcomes, and D) symptom models for physical outcomes.

A) Mental Component Score



B) Physical Component Score

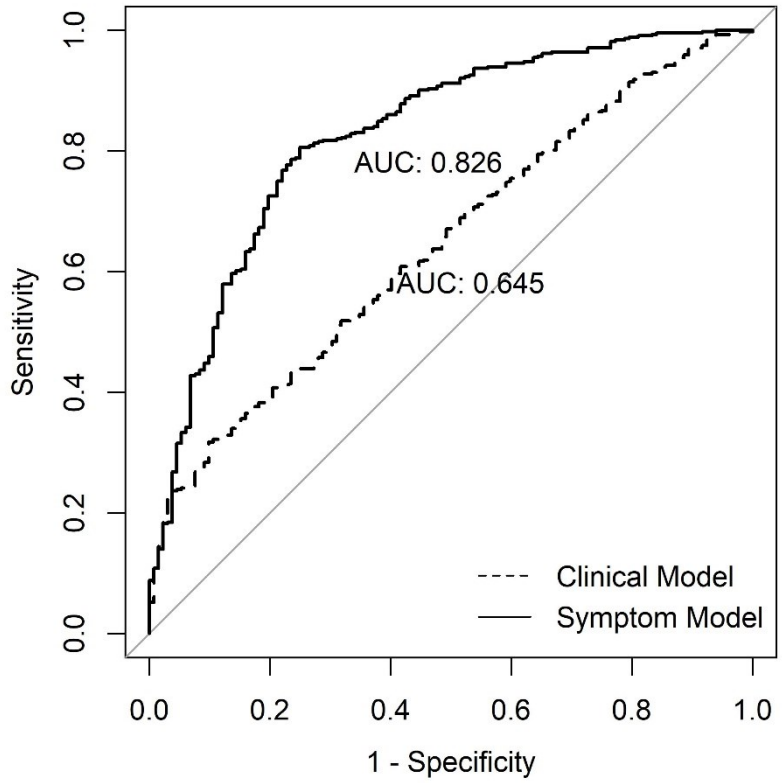


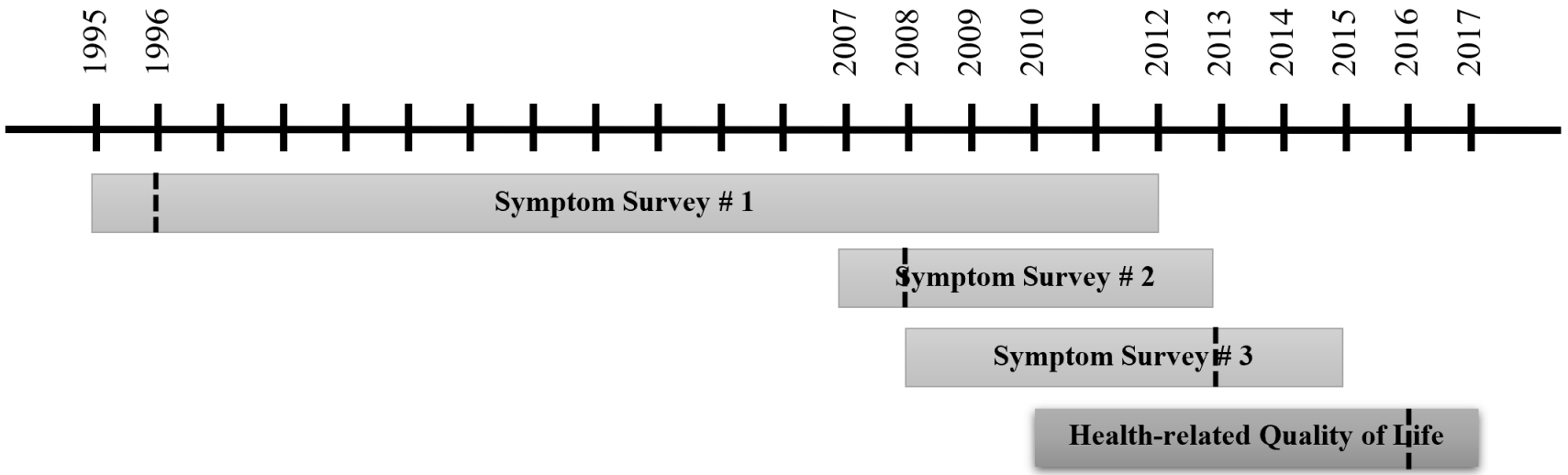
Figure 3.2: ROC Curves for HRQoL Models.
Receiver Operating Characteristic (ROC) Curves for A) Mental Component Score and B) Physical Component Score of SF-36 for 576 childhood cancer survivors comparing models without symptom patterns (clinical model, dashed line) and with symptom patterns (symptom model, solid line). Risk factors for the clinical model are selected from 35 clinical risk factors. Risk factors for the symptom model are selected from 35 clinical + 480 longitudinal symptom pattern risk factors. The ROC curves and the Area Under the ROC Curve (AUC) values are based on a cutoff of 40, representing clinically meaningful impairment.

3.5. Supplementary Information

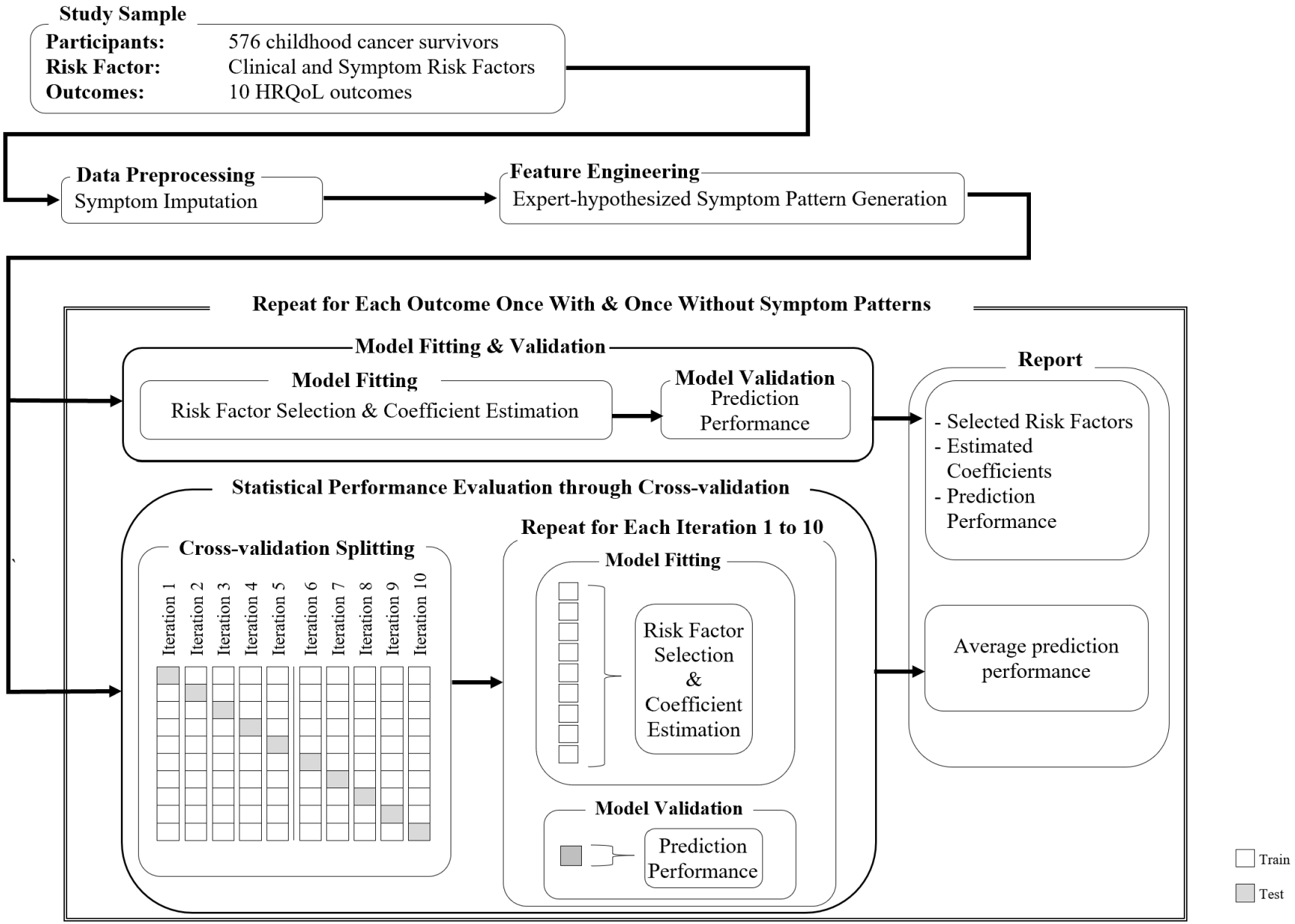
Supplementary Table 3.1. Performance evaluation for the 10 outcomes modelled without symptom patterns (clinical model) and with symptom patterns (symptom model).

	Clinical Model		Symptom Model		P-value**
	Model AUC*†	**Cross-validated AUC*‡**	**Model AUC*†**	**Cross-validated AUC*‡**	
Mental Component Score	0.571	0.559 (0.445, 0.730)	0.808	0.743 (0.626, 0.909)	<0.001
Mental Health	0.556	0.558 (0.395, 0.710)	0.809	0.739 (0.582, 0.909)	<0.001
Emotional Role Limitation	0.642	0.612 (0.435, 0.800)	0.854	0.800 (0.715, 0.859)	<0.001
Social Functioning	0.577	0.559 (0.442, 0.655)	0.785	0.752 (0.594, 0.885)	<0.001
Vitality	0.577	0.541 (0.441, 0.675)	0.745	0.696 (0.448, 0.854)	<0.001
Physical Component Score	0.645	0.626 (0.547, 0.760)	0.826	0.797 (0.696, 0.950)	<0.001
General Health Perception	0.632	0.583 (0.482, 0.672)	0.826	0.792 (0.709, 0.856)	<0.001
Physical Role Limitation	0.654	0.608 (0.509, 0.790)	0.807	0.781 (0.672, 0.906)	<0.001
Physical Functioning	0.661	0.650 (0.522, 0.815)	0.836	0.801 (0.671, 0.847)	<0.001
Bodily Pain	0.585	0.566 (0.441, 0.706)	0.794	0.767 (0.644, 0.944)	<0.001

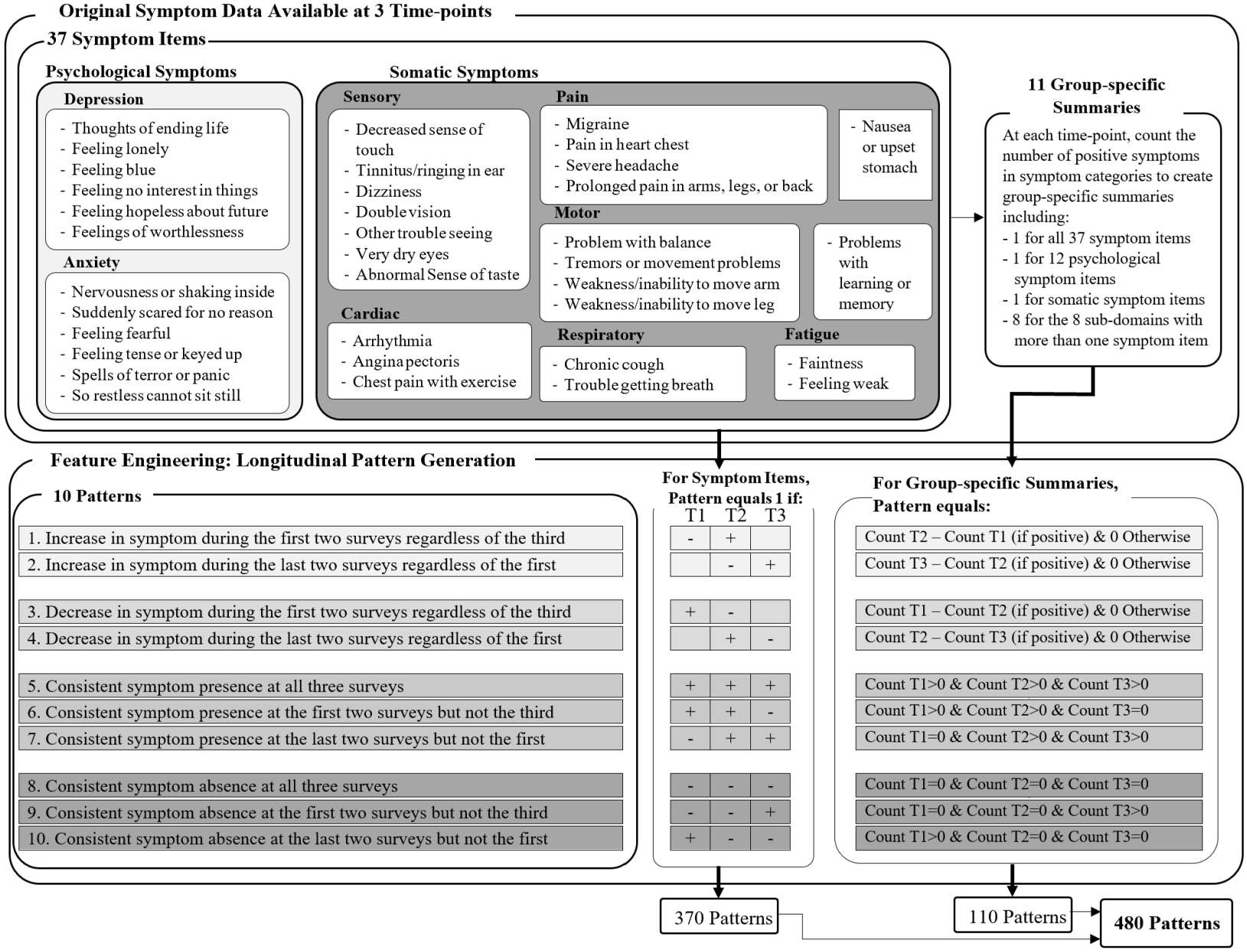
Note: *AUC: Area Under the Receiver Operating Characteristic Curve (AUC), calculated using a cutoff of 40 to indicate clinically meaningful impairment; † Model AUC represents the AUC value obtained from fitting and validating the model in the entire study sample; § Cross-validated AUC mean (range) is the average AUC value (along with the range) obtained from the 10 cross-validation iterations; ‡ P-value is reported based on DeLong's test comparing the two Receiver Operating Characteristic curves obtained from the entire study sample, one without the symptom data (clinical model) and one with the symptom data (symptom model).



Supplementary Figure 3.1. The range and median (represented by a dashed line) of completon times for the three symptom surveys and the Health-related Quality of Life Survey (Outcome Survey). The outcome survey for each individual survivor was assessed subsequent to the final symptom survey to prevent any temporal overlap between predictors and outcomes.



Supplementary Figure 3.2. Analysis Design Overview.



Supplementary Figure 3.3. Visual representation of feature engineering process.

CHAPTER 4

Predicting Personalized Burden of Multiple/Recurrent Health Conditions across the Lifespan in Childhood Cancer Survivors

4.1. Introduction

Across an individual's lifetime, a diverse array of health-related conditions may emerge, with some potentially occurring more than once. The Mean Cumulative Count (MCC) of multiple recurrent health conditions has proven to be a valuable measure, offering insights into the true average burden of disease in a population that extend beyond the cumulative incidence of single health conditions. The ability to predict a personalized, age-dependent estimation of this burden metric, with regards to individual-specific characteristics, is critical in facilitating timely tailored interventions and optimizing treatment protocols appropriate for each individual at the time of treatment decisions. This would empower individual patients and their care providers to utilize personalized predictions when deliberating strategies for preventing and managing health-related burden. Such personalized burden metric is particularly important when studying cohorts consisting of individuals with diverse treatment experiences and at an increased risk of developing a variety of conditions and premature mortality compared to the general population [5, 6].

In this paper, we sought to establish a framework for formulating a predictive model for age-specific, multitype/recurrent health conditions to help quantify expected health-related burden at the individual level. Our framework involves: (1) estimating hazard ratios and baseline hazard for each specific recurrent health condition individually and predicting risk scores for developing the condition at each small time interval over the lifespan of interest; (2) estimating mortality hazard ratios and baseline hazard, and predicting survival probabilities over the entire lifespan of interest;

(3) predicting accumulated risk of each specific recurrent condition separately over the entire lifespan of interest; (4) predicting cumulative count of each recurrent health condition to yield condition-specific PCC utilizing the information from steps 1-3; and (5) a summation of condition-specific PCCs obtained in step 4 over all conditions of interest to yield the overall PCC. The overall PCC shows the marginal age-specific provision of expected cumulative count/burden of multitype/recurrent health conditions given individual`s specific characteristics. Our framework acknowledges (1) competing risk of mortality, i.e., the fact that the succession of recurrent conditions may be terminated by death, (2) the possibility of recurrent episodes of several distinct conditions to count towards burden, and (3) the impact of various demographic and treatment variables on each condition and mortality, with mortality further being influenced by individual's experience of health conditions.

To illustrate the application of our proposed framework, we utilized data from childhood cancer survivors in the St. Jude Lifetime Cohort Study (SJLIFE) and St. Jude Long-Term Follow-Up Study (SJLTFU). Our aim was to predict the individual-level lifelong burden of cardiovascular recurrent chronic health conditions (CHCs).

The population of childhood cancer survivors is steadily increasing due to remarkable improvements in childhood cancer survival rates over the last several decades owing to the advancement in childhood cancer therapies[86, 87]. These long-term survivors particularly face an increased susceptibility to various CHCs in adulthood [36, 88-105], arising from diverse therapeutic exposures initiated during their cancer therapy [89-94, 96, 98-106]. These late-effects, and correspondingly the strategies for their screening and managing, are highly individualized, contingent upon factors such as the type and dose of treatment exposures [90, 91, 95, 97, 98, 104], age at the initial cancer diagnosis [104, 106], the time elapsed since treatment exposure [91], as

well as other common CHC-specific risk factors such as socio-demographic [95, 104], lifestyle [98] and clinical characteristics [95, 106]. This distinctive context provides an ideal case study to exemplify the value of our personalized burden metric.

This paper aims to predict the cumulative count of multitype/recurrent health conditions over time, considering competing-risk event of mortality that depends on the history of health conditions. The personalized approach adopted in this study holds the potential to uncover novel insights that may contribute to enhancing the long-term well-being of childhood cancer survivors. Utilizing clinically assessed CHC data collected longitudinally from the SJLIFE, this investigation endeavours to predict each CHC and mortality based on their respective predictors, which when combined, could provide a personalized quantification of morbidity with regards to multitype/recurrent CHCs.

4.2. Methods

4.2.1. Data source and study population

The SJLIFE study, initiated in 2007, is a retrospectively-constructed cohort study, with prospective follow-up, collecting data from ≥5-year survivors of childhood cancer. This cohort consisted of individuals who had received treatment or follow-up care for childhood cancer at the St. Jude Children's Research Hospital (SJCRH) and survived at least 5 years post cancer diagnosis. The SJLIFE study also has a community control group. The SJLIFE was designed to facilitate a longitudinal comprehensive clinical evaluation of the health outcomes among a lifetime cohort of adult survivors of childhood cancers. All participants in SJLIFE study underwent assessment for chronic health conditions using a standardized evaluation protocol at SJCRH. The study also recorded the dates of onset of these CHCs, tracked the longitudinal evolutions of disease, and collected treatment information, including cumulative dose-specific exposure to chemotherapy, surgical procedures, and radiation therapy. The detailed study design, eligibility criteria and validation method of medical events for SJLIFE have been previously reported elsewhere [81].

Initiated in the year 2000, SJLTFU aims to gather treatment, outcome, and late toxicity data for all SJLIFE eligible patients, encompassing those who did not participate in SJLIFE. The methodologies employed for abstracting treatment exposure data in SJLTFU mirror those utilized in SJLIFE [107]. Given their non-participation in SJLIFE, the chronic health condition outcomes of these individuals were not subject to direct clinical assessment.

The current study included a total of 4,336 participants (1,737 SJLTFU participants from SJLTFU and 2,599 from SJLIFE participants including 159 community controls who did not experience childhood cancer). Written informed consent was obtained from all participants before initiation

of any of the study procedures. The study protocol received approval from the SJCRH Institutional Review Board, and the ethical approval was obtained for the data analyses in this study.

4.2.2. Chronic health conditions (CHCs)

Individual CHCs and their grading

A total of 168 CHCs, originally clinically assessed within the framework of SJLIFE, were graded for severity as mild (grade 1), moderate (grade 2), severe/disabling (grade 3), life-threatening (grade 4) or death (grade 5) using the SJCRH-modified National Cancer Institute's Common Terminology Criteria for Adverse Events (CTCAE) [108, 109]. This modified CTCAE aimed to better encompass the spectrum of CHCs experienced by long-term survivors of childhood cancer.

Grouping CHCs

The 168 CHCs were grouped into 32 broad categories based on their shared clinical characteristics. This grouping was intended to allow for group-specific burden predictions. We implement all our calculations, such as hazard ratio or burden estimation, at the group level and hereafter we refer to these grouped conditions as conditions/CHCs.

4.2.3. Potential predictors

This study incorporated demographic characteristics, including sex (female/male), race (white/ non-white), and age at diagnosis of primary cancer. Furthermore, the study included three treatment variables including cumulative chemotherapy dose for anthracycline and mean dose of radiation therapy exposure for the heart and the brain, known to be associated with CHCs under investigation in this work. Information pertaining to original cancer diagnosis and detailed treatment protocols, including cumulative doses of chemotherapeutic agents, was abstracted by trained research staff from the primary medical records available at the treating institution/s using

a structured protocol. Additionally, primary radiation prescription records were utilized to estimate radiation dosimetry by radiation physicists [81].

4.2.4. The proposed metric and practical implication

This section introduces a framework designed to estimate a personalized health-related burden metric, referred to as PCC. This metric aims to quantify the marginal, age-specific cumulative count of multitype/recurrent health-related conditions over an individual`s lifetime, considering their unique characteristics. Here, we explain our framework through its five main steps and then apply it to our case study which seeks to quantify the burden of cardiovascular CHCs (5 CHC groups out of the 32 CHC groups in this study) in long-term childhood cancer survivors.

Step 1: Estimating hazard ratios of each recurrent health condition and predicting risk scores for developing the condition at each time interval over the lifespan of interest

Assuming a multiplicative effects for predictors, the rate function for condition *j* for an individual *i* at time *t* can be expressed as:

$$\rho_{ij}(t) = \rho_{0j}(t)\exp(x_{ij}^T\alpha_j) \quad (1)$$

Here, $\rho_{0j}(t)$ denotes the common baseline rate for condition *j* shared by all individuals, $x_{ij} = \left(x_{ij1}, \dots, x_{ijP_j}\right)^T$ is a P_j-dimensional vector of predictors where P_j is the number of predictors considered for condition *j*, and α_j is the corresponding P_j-dimensional vector of regression coefficients. Note that, the dimension of the coefficient vectors could be different by condition group based on variables known to affect the conditions of that group.

Following this formulation, estimating hazard ratios of recurrent conditions and the common baseline hazard is straightforward using conventional statistical tools such as Cox regression. Predictors for each condition could be selected based on prior knowledge or, alternatively, a

regularized regression approach can be employed to allow for considering the entire pool of potential predictors. After obtaining hazard ratios and baseline hazard for each condition, the risk of developing the condition at each small time-interval across the lifespan of interest for any individual could easily be predicted, given that the individual is alive at the beginning of the interval.

Step 2: Estimating mortality hazard ratios and predicting survival probabilities over the lifespan of interest

Mortality serves as the competing-risk event preluding the first-occurrence or recurrence of health conditions and is particularly important to consider when estimating marginal count of health conditions in high-risk populations. The count of health conditions is different in two individuals with similar rates of the condition but different mortality rates. Therefore, to estimate marginal condition counts accurately, it is essential to account for mortality.

Given that mortality may be strongly influenced by an individual's experience of multitype/recurrent health conditions, our approach to modelling mortality allows for including the cumulative count of recurrent health conditions as time-varying covariates, along with other time-invariant predictors such as demographic and treatment. The rate function for mortality for an individual *i* at time *t* can thus be given as:

$$\gamma_i(t) = \gamma_0(t) \exp\left\{ z_i^T \beta_1 + \sum_{j=1}^{J} \beta_{2j} N_{ij}(t) \right\}, \quad (2)$$

Here, $z_i = \left(z_{i1}, \ldots, z_{iQ}\right)^T$ is a Q-dimensional vector of covariates where Q is the number of time-invariant covariates considered for mortality, $\beta_1$is the corresponding Q-dimensional vector of regression coefficients, $N_{ij}(t)$ denotes the cumulative count of j^{th} recurrent health condition until

time t for individual i, and β_{2j} is the scaler regression coefficient corresponding to cumulative count of j^{th} recurrent condition.

One challenge in the above formulation is that the data for the recurrent health conditions is only available at the observed time-points, whereas we require this information for an individual's entire lifespan of interest. To overcome this challenge, these values could be imputed by extrapolating based on the last observed value or they could be predicted using the information obtained in step 1.

After obtaining cumulative count predictors for the entire lifespan of interest, estimating mortality hazard ratios and baseline hazard is a straightforward procedure using conventional statistical methods such as Cox regression. Once mortality ratios and baseline rate are obtained, survival probabilities can then easily be obtained for the lifespan of interest of a given individual.

Step 3: Predicting accumulated risk of each recurrent condition over the lifespan of interest

For each recurrent health condition, the accumulated risk by time t is simply predicted for the lifespan of the individual by summing up the corresponding hazards (instantaneous risks) for that condition up to time t.

Step 4: Predicting cumulative count of recurrent health condition to obtain condition-specific PCC

To predict the cumulative count of the recurrent health condition over a lifespan of interest, we first predict the expected count at all small intervals of time in the total lifespan of interest, then sum these values up to time t to get the cumulative count by time t. To obtain the aforementioned expected count at each small interval of time for a condition of interest, we multiply three values at the time interval including (1) survival probability at the beginning of that small interval of time which is obtainable through step 2, (2) risk for developing the condition during that small intervals

of time which is obtainable through step 1, and (3) the accumulated risk of developing the condition prior to that small interval of time which is obtainable through step 3. Hazard ratios of conditions and accumulated risk, obtained in step 1 and 3, alone cannot quantify the marginal count of conditions because the at-risk period for these hazard ratios is interrupted by death. Therefore, the information obtained in step 2 is also required.

Step 5: Summing condition-specific PCCs to obtain the overall PCC for the condition set of interest

Finally, the overall PCC considering multitype conditions is simply obtained by calculating the condition-specific PCCs for all conditions of interest and simply summing them.

4.2.5. Analytic Software

All analyses were performed using R version 4.21 (R Project for Statistical Computing).

4.3. Analysis of the Case Study by PCC Framework

To provide a detailed illustration of our metric, we employed PCC framework to predict the expected total count of cardiovascular CHCs in long-term childhood cancer survivors. Out of 32 broader groups of CHCs, five cardiovascular CHC groups (CHC1: arrhythmias, CHC2: cardiovascular dysfunction, CHC3: myocardial infarction, CHC4: stroke, CHC5: structural heart defects) were considered to illustrate the metric in detail. Supplementary table 4.1 shows the individuals CHCs contained in each of the five cardiovascular CHC groups. For each condition, we considered event as Common Terminology Criteria for Adverse Events (CTCAE) [109] grade 2 or above (i.e., moderate, severe/disabling, life-threatening or death) as event.

We formulated models of Equation (1) separately for each of the five CHC groups. In all five models, we incorporated key demographic predictors of age (time axis t), sex, race, and age at diagnosis. Additionally, each CHC-specific model included relevant treatments associated with that particular CHC group. Specifically, heart radiation and anthracycline dose were considered for CHC1, CHC2, CHC3, and CHC5, while brain radiation was included for CHC4. Furthermore, each CHC model considered potential inclusion of the interaction terms between its corresponding treatment predictors and age at diagnosis. For estimating these questions, the approach of group lasso with overlaps was used to allow for investigating and selecting the interaction terms only in the presence of their main effects [110, 111]. Group lasso, capable of retaining or removing the members of a specified group of predictors together, is particularly appealing for interaction investigation. To avoid overfitting while selecting predictors, we utilized 10-fold cross-validation. To calculate the 95% confidence intervals, we employed the 200-times bootstrap percentile method. Supplementary Table 4.2 shows the predictors used for modelling equation 1 for each of the cardiovascular CHC groups.

Models of Equation (2) were built starting with a pool of 32 predictors each representing the cumulative count of one of the 32 CHC groups, utilizing lasso to allow for selecting appropriate specific CHC groups affecting mortality. The 32 CHC groups considered in this study are listed in Supplementary Table 4.3. Like Equation (1), we applied a 10-fold cross-validation technique to prevent overfitting, and, for the computation of 95% confidence intervals, we utilized the bootstrap percentile method iterated 200 times.

4.4. Results

Characteristics of two example profiles utilized to showcase our framework are listed in Table 4.1, with some modifications in the values to ensure confidentiality. Table 4.2 provides a descriptive summary of the five cardiovascular CHCs among the childhood cancer survivors in this study, presenting the recurrence frequency for each CHC group. Starting with the most prevalent, the numbers (percentage) of survivors who experienced each cardiovascular CHC groups at least once were as follows: 454 (10.5%) for cardiovascular dysfunction; 273 (6.3%) for arrhythmias; 259 (6.0%) for myocardial infarction; 193 (4.5%) for structural heart defects; 166 (3.8%) for stroke.

The estimated hazard ratios for Equations (1) and (2) obtained through (group) lasso, along with their confidence intervals obtained via bootstrap resampling, are displayed in Tables 4.3 and 4.4. As indicated in Table 4.3, being a male survivor, younger age at diagnosis, and higher dose of radiation and chemotherapy were almost consistently identified to increase the risks of different cardiovascular events/recurrences in our analyses. More precisely, at a significance level of 0.05, being male was statistically significantly associated with a higher risk of cardiovascular dysfunction (hazard ratio = 1.55, 95% CI 1.22-1.95) and myocardial infarction (hazard ratio = 2.46, 95% CI 1.72-3.52). A one-year increase in age at the diagnosis of primary cancer was statistically significantly associated with a lower risk of stroke (hazard ratio = 0.95, 95% CI 0.92-0.99). Heart radiation dosage was identified to increase the risk of all cardiovascular CHCs except for stroke (the hazard ratios ranged from 1.33 to 1.98 per 10Gy of radiation), with the effect depending on the age at first diagnosis for myocardial infarction. In the case of stroke, brain radiation dosage, instead of heart radiation, was identified as increasing the risk with a hazard ratio of 1.27 per 10Gy of radiation (95% CI 1.16-1.38). As for chemotherapeutic agents, anthracycline dose was identified to increase the risk of cardiovascular dysfunction (hazard ratio = 1.40 per

100mg/m^2, 95% CI 1.25-1.57) and myocardial infarction (hazard ratios = 1.26). As shown in Table 4.4, six CHCs has been selected by lasso to predict mortality: survivors who experienced cardiovascular dysfunction, myocardial infarction, secondary and recurrent malignancies, and structural heart defects faced a higher risk of mortality, with estimated hazard ratios ranging from 1.18 to 1.93.

Table 4.5 presents the values of the cumulative count of CHC groups for the two randomly selected survivors: these values were observed up to the survivors` current ages but the values for time intervals beyond that point have been imputed by extrapolating from the last recorded data.

For the two selected survivor profiles, Figure 4.1 displays the predicted PCC curves (profile characteristics in Tables 4.1 and 4.5), assuming they will not die within the plotted time span, i.e., ages 20-60. For profile 1, the curves indicate that, by the age of 60, there would be an average PCC of 0.28 for arrhythmias, 0.35 for cardiovascular dysfunction, 0.11 for myocardial infarction, 0.27 for stroke, and 0.12 for structural heart defects, resulting in a total of 1.13 cardiovascular PCC. In other words, it is expected that an individual with profile 1 would experience an average of 1.13 cardiovascular CHCs by age 60.

For profile 2, by age 60, the predicted average PCC is 1.54 for arrhythmias, 2.14 for cardiovascular dysfunction, 0.78 for myocardial infarction, 0.04 for stroke, and 1.44 for structural heart defects, with a total of 5.94 cardiovascular PCC. One might compare the values of CHC count from Table 4.5 and Figure 4.1.

4.5. Discussion

We have proposed a novel framework for predicting the expected burden of the health-related conditions at the individual-level. This framework accounts for the recurrence and coexistence of multiple conditions in burden prediction and integrates an individual`s unique constellation of risk factors. Applying PCC to the data of long-term childhood cancer survivors, who are at risk of multiple recurrences of different cardiovascular CHCs, has shown its potential to predict the marginal count of each CHC by a given age for an individual given their unique characteristics. The framework developed in this work could have a wide-ranging applicability beyond CHC events for childhood cancer survivors and is also applicable to any population who may be at risk for multitype/recurrent health events.

The PCC approach offers improved quantification of the true burden, better statistical precision and statistical power, compared to a time-to-first event analysis focusing on the first-occurrence of a single condition. First, not confining the burden index to a single health-related condition could help better represent the overall disease burden experienced by high-risk individuals at risk of several conditions. Similarly, taking into account recurrences of conditions, rather than only the first event which cumulative incidence is concerned, is better reflective of the burden faced by individuals who might repeatedly experience the same condition. Second, PCC improves the statistical precision and enhances statistical power by not truncating the analysis at the occurrence of the first-event as in time-to-first event analysis. Specifically, PCC continues to follow all individuals even after their first event. This contrasts with time-to-first-event analysis where hazard ratios tend to underestimate the true value as high-risk patients experience events early, while the low-risk individuals will remain under observation longer [112]. Third, PCC leads to a more precise burden value as it employs an individualized cumulative burden approach compared

to the conventional recurrent event analysis. While the conventional recurrent-event analysis considers the tendency of a first event to cause a subsequent increase in risk for the next event, it uses a non-individualized cumulative burden approach. PCC, on the other hand, incorporated individual variations in risk factors to estimate the cumulative burden by employing the technique of regression analysis to model all conditions and mortality, integrating subject-specific information such as demographic and treatment covariates. Finally, PCC further incorporates the fact that an individual cannot experience any further recurrent event once the terminal event, death, has been experienced.

Our proposed metric, the PCC, is based on the same rationale as the commonly used approach in the analysis of time-to-first-event to predict cumulative incidence (probability) from cause-specific hazard ratios and survival probability in the presence of competing-risk event [113-115]. Both procedures are rooted in the fact that rates of condition first-occurrence/recurrence only reflect what happens locally in time among individuals at risk (i.e., instantaneous risk conditional on survival), and thus cannot provide any marginal values without accounting for competing-risk event of mortality. The marginal expected count of recurrent events is vulnerable to survival probability on the ground that the number of recurrent events increases when one lives longer [116]. Most importantly, the hazard ratios of recurrent events obtained in step 1, the survival probabilities obtained in step 2, and the accumulated risk of recurrent events obtained in step 3 jointly characterize the participant's count of recurrent event, warranting the PCC to be used to check the influence of a new treatment by including it in modelling procedures in step 1 and 2. Furthermore, our approach leverages familiar regression tools, and thus, is conductible by all statistical packages that could run time-to-event regression. Additionally, utilizing lasso facilitated our method's dealing with a large number of potential predictors [110, 111].

In summary, the PCC burden metric predicts the number of recurrent events of multiple types of conditions that may develop in an individual during a specified time period, taking into account the condition- and mortality-specific predictors and the competing-risk event of mortality. This burden prediction enables researchers to investigate therapeutic interventions, predict individual-level burdens of multitype/recurrent health conditions, and assists in tailoring treatment options by consulting the predicted burden of various treatment options.

Table 4.1: Characteristics of the two selected example profiles.

	Example Profile 1	**Example Profile 2**
Sex (Male)	No	Yes
Race (White)	No	Yes
Age at diagnosis of primary cancer [Years]	1.58	14.9
Current age [Years]	44.7	49.5
Mean dose of chemotherapy		
Anthracycline [DOXED 100mg/m^2]	1.48	4.51
Mean dose of radiation therapy [10Gy]		
Heart	0.04	4.65
Brain	2	0

Table 4.2: Frequency of recurrence for each cardiovascular CHC among study participants.

	Frequency of recurrence			
Cardiovascular CHC	0	1	2	≥3
CHC1: Arrhythmias	4,063	201	40	32
CHC2: Cardiovascular dysfunction	3,882	406	35	13
CHC3: Myocardial infarction	4,077	218	32	9
CHC4: Stroke	4,170	125	27	14
CHC5: Structural heart defects	4,143	150	31	12

Note: CHC-Chronic Health Condition

Table 4.3: Estimated hazard ratios for predictors of cardiovascular CHC groups.

	CHC1: Arrhythmias		CHC2: Cardiovascular dysfunction		CHC3: Myocardial infarction		CHC4: Stroke		CHC5: Structural heart defects	
	HR	95% CI	HR	95% CI	HR	95% CI	HR	95% CI	HR	95% CI
Sex (Male)	0.86	0.63, 1.19	1.55**	1.22, 1.95	2.46**	1.72, 3.52	1.34	0.89, 2.01	1.11	0.76, 1.61
Race (White)	1.21	0.77, 1.90	0.8	0.60, 1.08	1.19	0.79, 1.78	0.68	0.43, 1.08	1.02	0.64, 1.62
Age at diagnosis [Years]	0.97	0.94, 1.01	0.98	0.95, 1.00	1.02	0.97, 1.07	0.95*	0.92, 0.99	0.96	0.91, 1.00
Heart radiation [10Gy]	1.33*	1.06, 1.65	1.38**	1.28, 1.48	1.98**	1.61, 2.43			1.92**	1.54, 2.39
Brain radiation [10Gy]							1.27**	1.16, 1.38		
Anthracycline dose [DOXED 100mg/m^2]	1.17	0.98, 1.39	1.40**	1.25, 1.57	1.26*	1.03, 1.53			1.10	0.98, 1.24
Age at diagnosis × Heart radiation	1.00	0.99, 1.02			0.98*	0.96, 0.99			1.00	0.98, 1.01
Age at diagnosis × Brain radiation							1.01	1.00, 1.02		
Age at diagnosis × Anthracycline dose	1.00	0.99, 1.02	1.00	0.99, 1.02	0.98	0.96, 1.00				

Note: HR-hazard ratio; CI-Confidence Interval; *P-value<0.05; **P-value<0.01

Table 4.4: Estimated hazard ratios for predictors of mortality.

	HR	95% CI
Cumulative number of recurrent events of CHC1: Arrhythmias	1.00	0.84, 1.19
Cumulative number of recurrent events of CHC2: Cardiovascular dysfunction	1.42	1.13, 1.78
Cumulative number of recurrent events of CHC3: Myocardial infarction	1.18	1.01, 1.38
Cumulative number of recurrent events of CHC5: Structural heart defects	1.35	1.07, 1.71
Cumulative number of recurrent events of CHC6: Kidney injury	1.61	1.00, 2.59
Cumulative number of recurrent events of CHC7: Secondary and Recurrent Malignancies	1.93	1.74, 2.13

Note: HR-hazard ratio; CI-Confidence Interval

Table 4.5: Recurrence of cardiovascular CHCs for two selected example profiles.

Age range	Cumulative Count of CHC Group					
	CHC1: Arrhythmias	CHC2: Cardiovascular dysfunction	CHC3: Myocardial infarction	CHC5: Structural heart defects	CHC6: Kidney injury	CHC7: Secondary and Recurrent Malignancies
Example Profile 1						
0 - <10	0	0	0	0	0	0
10 - <20	0	0	0	0	0	0
20 - <30	0	1	0	0	1	0
30 - <40	0	1	0	0	1	0
40 - <50	0	1	0	0	1	0
50 - <60	0	1	0	0	1	0
≥ 60	0	1	0	0	1	0
Example Profile 1						
0 - <10	0	0	0	0	0	0
10 - <20	0	0	0	0	0	0
20 - <30	0	1	0	0	0	0
30 - <40	0	2	0	0	1	0
40 - <50	1	2	0	0	1	1
50 - <60	1	2	0	0	1	1
≥ 60	1	2	0	0	1	1

A) Example profile 1

B) Example profile 2

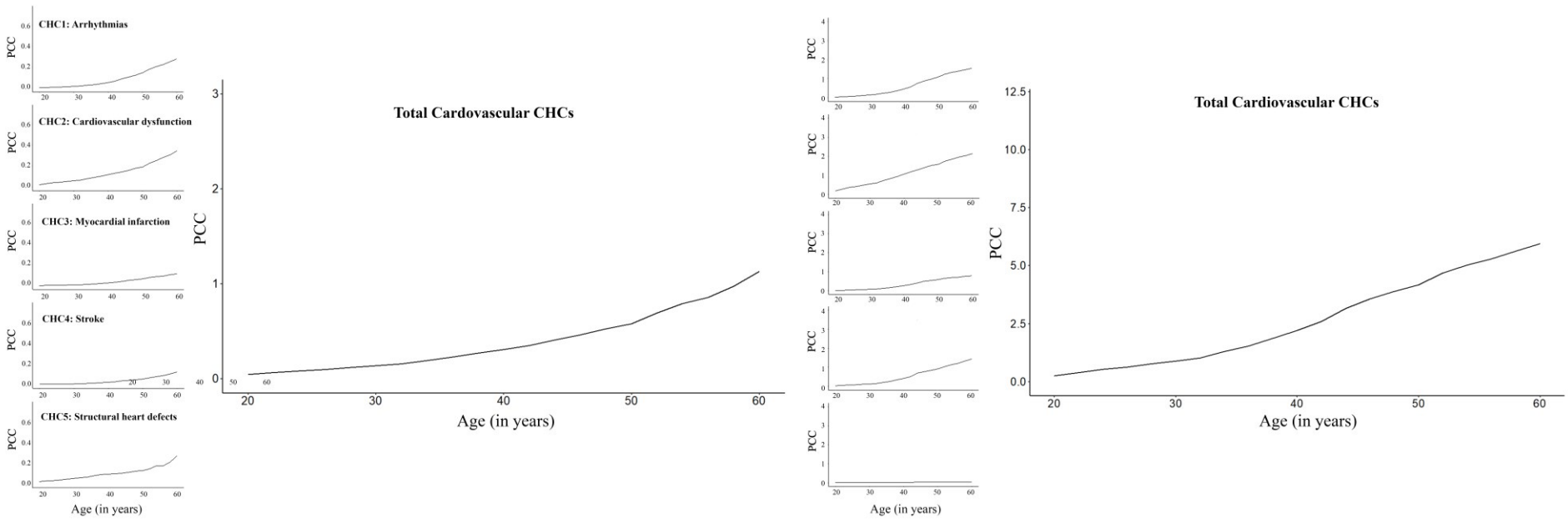


Figure 4.1: PCC curves for example profiles.

Estimated PCC curves for each of the five cardiovascular CHCs individually obtained in step 4 and an overall PCC curve for the total five cardiovascular PCCs obtained in step 5, shown separately for the two example profiles

4.6. Supplementary Information

Supplementary Table 4.1. Grouped and individual cardiovascular chronic health conditions

CHC1: Arrhythmias
- Atrioventricular heart block
- Conduction abnormalities
- Prolonged QT interval
- Cardiac dysrhythmia
- Sinus bradycardia
- Sinus tachycardia

CHC2: Cardiovascular dysfunction
- Cardiomyopathy
- Right ventricular systolic dysfunction
- Cor pulmonale
- Pulmonary hypertension

CHC3: Myocardial infarction
- Myocardial infarction

CHC4: Stroke
- Intercranial hemorrhage
- Cerebrovascular accident
- Cerebrovascular disease

CHC5: Structural heart defects
- Heart valve disorder
- Pericarditis
- Aortic root aneurysm
- Atrial myxoma

Supplementary Table 4.2. Predictors considered for each CHC group

CHC group	Demographic	Treatment	Interaction of treatment variables	Interaction of age at diagnosis with treatment variables
CHC1: arrhythmias	- Sex - Race - Age at diagnosis	- Heart radiation - Anthracycline dose	Heart radiation * Anthracycline dose	Age at diagnosis * Heart radiation Age at diagnosis * Anthracycline dose
CHC2: Cardiovascular dysfunction	- Sex - Race - Age at diagnosis	- Heart radiation - Anthracycline dose	- Heart radiation * Anthracycline dose	- Age at diagnosis * Heart radiation - Age at diagnosis * Anthracycline dose
CHC3: Myocardial infarction	- Sex - Race - Age at diagnosis	- Heart radiation - Anthracycline dose	- Heart radiation * Anthracycline dose	- Age at diagnosis * Heart radiation - Age at diagnosis * Anthracycline dose
CHC4: Stroke	- Sex - Race - Age at diagnosis	- Brain radiation		- Age at diagnosis * Brain radiation
CHC5: Structural heart defects	- Sex - Race - Age at diagnosis	- Heart radiation - Anthracycline dose	- Heart radiation * Anthracycline dose	- Age at diagnosis * Heart radiation - Age at diagnosis * Anthracycline dose

Supplementary Table 4.3. Grouped and individual cardiovascular chronic health conditions

CHC1:	Arrhythmias
CHC2:	Cardiovascular dysfunction
CHC3:	Myocardial infarction
CHC4:	Stroke
CHC5:	Structural heart defects
CHC6:	Kidney injury
CHC7:	Secondary and Recurrent Malignancies
CHC8:	Essential hypertension
CHC9:	Dyslipidemia
CHC10:	Obstructive respiratory disorder
CHC11:	Functional pulmonary deficit
CHC12:	Esophageal disorders
CHC13:	Hepatic disorders
CHC14:	Disorders of the gallbladder
CHC15:	Disorder of the female reproductive system
CHC16:	Disorder of the male reproductive system
CHC17:	Male or female hypogonadism
CHC18:	Obesity (by BMI)
CHC19:	Thyroid disorders
CHC20:	Abnormal glucose metabolism
CHC21:	Obstructive urinary disorder
CHC22:	Urinary dysfunction
CHC23:	Amputation
CHC24:	Joint disease
CHC25:	Peripheral nervous system disorder
CHC26:	Spine disorder
CHC27:	Peripheral nervous system disorder
CHC28:	Seizures
CHC29:	Severe headaches
CHC30:	Hearing Loss
CHC31:	Ocular disorders
CHC32:	Immunology and Infectious Systems

CHAPTER 5

DISCUSSION

5.1. Overview

In this dissertation, we undertook a multifaceted exploration of the complex conditions of childhood cancer survivorship. My work represents a concerted effort to transcend some of the conventional approaches typically employed in investigating the late-effects of childhood cancer and its treatment. This departure includes (1) moving beyond reliance on a priori hypothesized risk factors and one-element-at-a-time addition/removal of risk factors and to investigating diverse spectrum of survivors' risk factors collectively, (2) incorporating continuous patient observations over time to account for the longitudinal journey of survivors throughout their cancer care-survivorship continuum, (3) acknowledging survivors' perspectives on their health status, and (4) integrating comprehensive multi-dimensional/multitype morbidity measures for outcome assessment.

To achieve this, we leveraged exceptional data sources of CCSS [80] and SJLIFE [81] for the required measurements and utilized cutting-edge statistical/machine-learning techniques to facilitate the required analysis. Through these efforts, we pave the way for a more comprehensive and personalized approach to cancer treatment and survivor care. My approach not only holds the promise of discovering previously unknown factors affecting long-term survivors but also could offer improved precision in identifying individuals who could benefit most from early treatment/preventative interventions. This dissertation led to three distinct yet interconnected papers, all emphasizing the increased involvement of survivors and putting their whole story in perspective, and each shedding light on crucial areas of concern for childhood cancer survivors.

Our overarching goal in these papers is to improve the quality of life and well-being of survivors by deepening our understanding of the factors that shape their late-effects. In the upcoming sections of the discussion, we will revisit the factors emphasized to achieve the goals of this dissertation, specify the sections in which each was addressed, and highlight their significance within each respective section.

5.1.1. Assessing a broad array of potential risk factors collectively over and beyond individual factor or one-subset-at-a-time assessment for enhanced late-effect modelling

The prevailing methodology employed in late-effects studies often concentrates on evaluating a limited subset of previously hypothesized factors or, sometimes, utilize techniques that adopt a greedy assessment of the factors to suggest candidate combinations of risk factors, inadvertently overlooking the potential synergies that can arise from a novel combination of factors. Despite the widespread emphasis and understanding on utilizing prediction tools for personalized survivorship care [70, 71, 117] and the huge focus of machine learning society on designing tools for such work , there is a big gap between the two communities, and it takes a lot of time for a designed methodology to gain trust and be utilized in the health community. So, while this kind of innovative methodologies are new in survivorship analysis, there has been a lot of theoretical encouragement and invitations through the years [16]. The primary strength of the analyses presented in this thesis lies in the methodological frameworks employed to uncover combinations of factors that collectively explain/predict patient outcomes. Presently, while the significance of these matters is acknowledged, they remain somewhat underscored.

Assessing potential risk factor collectively received attention in all three works of this study. We utilized different versions of penalized regression aimed at efficiently identifying collections of

factors that exhibit enhanced collective performance. Specifically, we used elastic net, BIEN, lasso, and group lasso based on the needs in different sections for this dissertation.

The first study focused on designing a methodology capable of collective assessment of numerous potential risk factors, while the second and third works utilized methodologies with such capability to address real-world problems. The methodology in the first study, referred to as BIEN, was developed with the central aim of facilitating collective investigation of a broad pool of potential factors, accounting for the relationships among all of them. The existing method of elastic net was utilized to suggest candidate risk factor sets in our innovative BIEN methodology The rationale behind BIEN was that selection of relevant predictor and estimation of their effect sizes are interdependent and must be concurrently considered in the same optimization function. BIEN's concurrent investigation of both the choice of the covariates and their effect sizes could potentially explain its consistently similar or superior performance compared to the alternative methodologies under consideration. This, in turn, could potentially offer unique insights that may not be attainable through alternative approaches lacking this capability.

To collectively investigate a diverse set of risk factors, encompassing clinical and symptom variables, for modelling 10 HRQoL scores in childhood cancer survivors, the second study utilized BIEN, the methodology developed in the first work. The collective assessment is particularly appealing in this context of symptom variables due to their high collinearity. The selected risk factors in this work, along with the direction of their effects, aligned with expert expectations in the field, establishing confidence in our approach to addressing the problem. Furthermore, with many risk factors consistently selected for several HRQoL scores, additional confidence is lent to our approach to addressing the problem. In general, the high inclination toward selecting symptoms in this work, rather than the more traditionally assessed clinical variables, and the

appreciable improvement in outcome prediction associated with this choice, suggest enhanced insight attainable through the collection and analysis of diverse novel sets of variables and, consequently, a potential to provide contribution for survivorship care guidelines. Particularly, the selected symptoms in this work can guide survivors on when to seek clinic referrals and facilitate effective communication between clinicians and survivors during clinical visits, assisting clinicians in making well-informed referral decisions. In studies like this where we face a broad array of variables to select from, using advanced statistical/machine learning tools help extract timely knowledge from the rapidly increasing number of risk factors being collected.

To collectively investigate and select relevant risk factors and quantify their effect from the set of hypothesized risk factors for mortality and conditions of interest within the PCC framework, the third study employed existing lasso and group lasso methodologies, respectively, during the modelling process for each. Lasso was utilized for selecting relevant CHCs contributing to the risk of death out of the wide range of CHCs available in childhood cancer data. Groups lasso was used for modelling CHCs as it facilitates investigation of interaction terms between demographic and treatment variables as groups. In comprehensive assessments like these, employing automated procedures helps alleviate the workload/input on statistical analyst and field expert to make every decision. The PCC metric derived from this intricate procedure provides highly personalized insights into the burden of disease for an individual, taking into account their unique experiences. This personalized burden prediction approach has the potential to offer novel tailored insights for treatment decisions in cancer patients at the time of initial treatment, as well as for preventative and treatment intervention decisions during survivorship care.

5.1.2. Incorporating subsequent experience of cancer and its treatment through time over and beyond initial experience for enhanced late-effect modelling

Studies on late-effects in childhood cancer survivors often primarily focus on variables measured at the time of the initial cancer treatment with less emphasis on comprehending the impact of survivors' subsequent experiences throughout their journey post-initial treatment. This limits the exploration of the diverse, dynamic, and intensifying experiences of risk factors as survivors navigate the extended journey of cancer survivorship [72]. Despite the current widespread recognition of this important concept, there is still ample room for improvement to accomplish this endeavour. Both the second and third studies in this dissertation address this concept to account for experiences of survivors more adequately across their survivorship continuum, each employing a distinct approach selected based on the context. The insights presented in both pieces of work can be attributed, in part, to this approach that may not be discerned through considering only variables considered during the time of treatment. We hope that the approaches we used to incorporate the longitudinal journey of childhood cancer survivors in their modelling of late-effects provide suggestive hints for alternative avenues to implement this important concept in future research.

In the second study, our approach for incorporating the longitudinal symptom experience of survivors started with consolidating the symptom experience into a set of patterns. The set of patterns generated for each symptom is supposed to represent all possible potential fluctuations of that symptom over time. More specifically, we transformed the experience of each symptom available in three time-points following the initial treatments into new predictors, each representing a potential pattern of symptom at the three time-points. For example, if a specific pattern is present at all three assessment points, then, from the many patterns created, the one

representing consistent presence at three time-points receive a value of 1 and all other patterns receive a value of 0. The rationale behind this is to base the modelling on patterns, as more stable predictors compared to the symptom states themselves, to reduce the possibility of small variations in the data leading to a dramatic change in the selected model. The fact that symptom patterns appeared as powerful indicators of an individual's well-being in this work, overshadowing the at-the-time-of-treatment variables, is partially explained by the fact that they are representative of subsequent state of the health status. This could also suggest that symptoms are acting as mediators for the relationship of the HRQoL and treatment factors, leading to implications for symptom surveillance during follow-ups and interventions aimed at alleviating specific symptoms, ultimately enhancing the long-term well-being of survivors.

In the third study, the extended journey of survivors is captured in PCC prediction through two concepts. Firstly, recurrent event analysis was employed for CHC modelling instead of first event analysis, extending our calculations to account for the survivor experience beyond the initial occurrence of an event. Besides offering a more realistic consideration of risk, this approach also brings about enhanced statistical power. Secondly, acknowledging the heightened risk of various CHCs and the associated increased mortality risk over time for survivors, we derived the cumulative experience of CHCs as time-dependent measures of morbidity to be used as a predictor in the mortality model. These two considerations together pave the way toward more precise and personalized quantification of the burden survivors face, potentially leading to improved decision making for an individual person. The observed growing difference in life-time burdens over time between the two example survivors int this study, who had distinct demographic and treatment exposures, advocates for the importance of incorporating personalized longitudinal experiences into consideration.

5.1.3. Incorporating patient's voice over and beyond objective measures for enhanced late-effect modelling

In recent years, there has been a recognized focus on integrating survivors' voices into cancer survivorship research, aiming to illuminate and incorporate the highly individualized aspects of their experiences. This initiative is undertaken with a goal to deliver patient-centered survivorship care and make timely treatment/intervention decisions, even before a problem becomes clinically apparent [118]. While the importance of this crucial concept has been practiced through the use of Patient-Reported Outcomes (PROs), such as symptoms and HRQoL, there is still significant potential for more efficient and timely utilization of patients' voices, considering the limitations of the methodological approach employed. The second study in this work focused on addressing this aspect.

In its approach to leveraging symptoms, the data-driven method employed to identify imporatnt risk factors from a plethora of potential options in the third work holds particular promise within the context of PROs. This is because a hypothesis-driven approach, reliant on prior knowledge of underlying etiology, can be challenging to formulate promptly given the abundance of routinely collected PROs and their less stringent definition. While evidence on relevant PRO risk factors has been accumulating due to heightened focus in this area, such evidence may not always encompass the latest collected symptoms. We envision our approach serving as a pipeline for extracting knowledge in similar scenarios and aiding in the construction of user-friendly decision-making tools, such as risk calculators, to be incorporated into survivorship care plans or personal portals.

5.1.4. Incorporating comprehensive multi-dimensional/multitype measures of morbidity over and beyond single morbidity measures for enhanced late-effect modelling

In the late-effect modeling within this dissertation, we transcended singular measures of morbidity, moving from considering one specific health problem at a time to a more holistic approach, with the aim of comprehensively quantifying the magnitude of the burden faced by each survivor of childhood cancer given their characteristics. This critical concept has been a focal point in both the second and third studies within this dissertation, with the second work utilizing multi-dimensional HRQoL scores and the third work utilizing a variety of CHCs. This approach holds the potential to highlight nuanced differences in the magnitude of the burden faced by childhood cancer patients, a subtlety that may not be as apparent when concentrating on each individual health dimension in isolation. We anticipate that this multi-dimensional and comprehensive approach contributes additional insights for improving the overall health status of childhood cancer survivors, surpassing the insights derived from studies focusing on single morbidity measures.

5.2. Strengths and limitations

Collectively, the efforts in this study has several strengths and advantages: (1) identifying novel predictors of late-effects and gain nuanced insights into late-effect mechanisms, (2) quantifying the burden faced by survivors more accurately and realistically, (3) deriving knowledge from the incoming massive data at the real-time that they are being collected, and (4) better and earlier identification of the high-risk individuals as targets of interventions to avoid the possible delays in decisions.

Several limitations need to be recognized in relation to this study. Firstly, the fact that both construction and evaluation of the models in this study rely on a single source of data may constrain

the generalizability of the findings to other childhood cancer survivors. St. Jude Children's Research Hospital is the only National Cancer Institute accredited Cancer Center which focuses on childhood cancer. Furthermore, patients are treated with no cost to the family at St. Jude. As such, the experiences of childhood cancer survivors who had been treated at St. Jude may be different from those who are treated elsewhere. Ensuring our model's applicability in general requires evaluation in an independent population of survivors not involved in the construction of models. In future research, we must assess the robustness and generalizability of our findings utilizing different childhood cancer survivor cohorts. Related to this point is the unavailability of comparable cohorts in the world. While some European countries have survivor cohorts, many mimicking or using the same protocols as SJLIFE and CCSS, the size of the cohort and/or length and assessment of the follow up differ. The lack of comparable cohorts is a general challenging issue in this field.

Secondly, implementation and dissemination of the work is hardly addressed in our work thus far. A challenge linked to the utilization of the models developed in this work is that healthcare providers may need some understanding of the procedures and methodologies behind them to trust and apply them, and this could be challenging due to the incorporation of multiple tools in constructing the models. This "buy in" issue may especially be of concern since we have not had a robust replication of the results in an independent cohort. Usefulness of the findings for survivorship care must be established in different ways, given the lack of replication opportunity.

Third, symptoms change over time and healthcare providers may find it tedious or time consuming to collect symptom data longitudinally from each patient within the limited time in their practice. We did not involve implementation science perspectives in this research. Ideally, from the study design, conduct including data collection to the data analysis and evaluation of its results, and

dissemination/implementation of the findings must reflect the opinions and feedback from the various stakeholders. We admit that we did not go through this step in this dissertation research fully, and focused mainly on biostatistical perspectives of the issues. In the future, not only the robust replication of the findings, but also dissemination and implementation issues must be taken up and addressed. Regarding this point, it is particularly important to involve survivors in terms of what they need and care, rather than solely approaching the problem from researchers' perspectives. While we incorporated survivors' voices through PROs, our research as a whole must reflect the needs and preferences of survivors and address them through research activities.

Fourth, while technologies and tools exist to help the dissemination and implementation of our findings, we have not utilized them in our research thus far. For example, we have developed St. Jude Survivorship Portal on St. Jude Cloud, through which all cohort data of SJLIFE, including clinical and genetic data, are fully available for online access for anyone in the world. This public sharing of the data is consistent with the nationwide efforts by the National Cancer Institute for making all childhood cancer data available on a single platform so that researchers can fully utilize individual institutions data jointly. The St. Jude Survivorship Portal is equipped with cutting-edge statistical analysis and visualization capabilities through implementation and utilization of R so that users can not only access but explore and conduct statistical analysis, including advanced regression analyses, of the SJLIFE data: no other online data sharing tool has these capacities that share whole genome sequencing data as well as clinically-assessed health condition data and comprehensive cancer treatment data including radiation dosimetry and cumulative doses of various chemotherapies. Our plan is to implement the third paper of my thesis, the personalized prediction models of chronic health conditions, as part of the St. Jude Survivorship Portal.

Finally, statistically, the methodology presented in our initial paper, BIEN, has certain limitations in accurately quantifying the uncertainty linked to estimated coefficients, including confidence intervals and p-values, or any other pertinent concepts in statistical inference. Such measures of uncertainty used traditionally in statistics assume that the predictors of the model have been selected a priori and are not data-driven, and, thus, would lead to overoptimistic results in the context of data-driven selection of covariates. The challenge of uncertainty quantification related to variable selection, a problem referred to as selective inference [119], remains a subject for future exploration. Additionally, while we focused on incorporating components familiar to health researchers into our methodology in order to boost its acceptance within the health research community, it is important to recognize that BIEN may still present challenges for those seeking a thorough understanding of theory prior to application, as it demands a more substantial background in theoretical knowledge. Lastly, since our primary emphasis lies in leveraging empirical results to advocate for our methodology, conducting more comprehensive simulations could further enhance the credibility of our methodology. This includes exploring alternative scenarios with various number of covariates and sample sizes, as well as scenarios featuring less multicollinearity compared to the present study,

5.3. Conclusions and clinical/public health implications

The three papers together emphasized the multi-faceted nature of the experience of childhood cancer survivors, holding important implications for understanding of childhood cancer survivors` late-effects. Paper 1 introduced a novel methodology to explore the collective impact of numerous potential covariates, emphasizing consideration of the entire narrative of the covariates, rather than fragmented parts. Paper 2 attempted to contribute to the existing body of knowledge on late-effects of childhood cancer and its treatment by placing emphasis on patients' voices and the longitudinal

journey of survivors subsequent to the initial cancer experience. Paper 3 sought to enhance the current understanding of the late-effects of childhood cancer by utilizing a framework that placed emphasize on the longitudinal experience of survivors and utilizing relevant risk factors in all components being modelled, and accounting for interruption of the at-risk period by death.

In our endeavor to achieve the objectives outlined in this dissertation, we designed and implemented cutting-edge statistical/machine learning methodologies, enabling us to harness diverse data sources. This dissertation serves as an effort to respond to the encouragement and calls by researchers in the field of cancer survivorship research in recent years, advocating for the utilization of advanced methodologies for knowledge discovery and enhanced precision in personalized survivorship care. Beyond contributing to a deeper understanding of survivors' health states, we hope that the innovative data-driven approaches, frameworks, and pipelines in this dissertation serve as valuable tools for future research in this field and other related fields and inspire further developments and applications.

REFERENCES

1. Atun, R., et al., *Sustainable care for children with cancer: a Lancet Oncology Commission.* The Lancet Oncology, 2020. **21**(4): p. e185-e224.
2. Mariotto, A.B., et al., *Long-term survivors of childhood cancers in the United States.* Cancer Epidemiology, Biomarkers & Prevention, 2009. **18**(4): p. 1033-1040.
3. Robison, L.L. and M.M. Hudson, *Survivors of childhood and adolescent cancer: life-long risks and responsibilities.* Nature Reviews Cancer, 2014. **14**(1): p. 61-70.
4. Siegel, R.L., et al., *Cancer statistics, 2023.* CA: a cancer journal for clinicians, 2023. **73**(1): p. 17-48.
5. Bhatia, S. and L.L. Robison, *Cancer survivorship research: opportunities and future needs for expanding the research base*. 2008, AACR. p. 1551-1557.
6. Bhakta, N., et al., *The cumulative burden of surviving childhood cancer: an initial report from the St Jude Lifetime Cohort Study (SJLIFE).* The Lancet, 2017. **390**(10112): p. 2569-2582.
7. Huang, I.-C., et al., *Association between the prevalence of symptoms and health-related quality of life in adult survivors of childhood cancer: a report from the St Jude Lifetime Cohort study.* Journal of Clinical Oncology, 2013. **31**(33): p. 4242.
8. Yeh, J.M., et al., *A model-based estimate of cumulative excess mortality in survivors of childhood cancer.* Annals of internal medicine, 2010. **152**(7): p. 409-417.
9. Williams, A.M., et al., *Rethinking success in pediatric oncology: beyond 5-year survival.* Journal of Clinical Oncology, 2021. **39**(20): p. 2227.
10. Yabroff, K.R., et al., *Economic burden of cancer in the United States: estimates, projections, and future research.* Cancer epidemiology, biomarkers & prevention, 2011. **20**(10): p. 2006-2014.
11. Given, B.A., P. Sherwood, and C.W. Given, *Support for caregivers of cancer patients: transition after active treatment.* Cancer epidemiology, biomarkers & prevention, 2011. **20**(10): p. 2015-2021.
12. Lancet, T., *Cancer care: beyond survival*. 2022. p. 1441.
13. Council, N.R., *Childhood cancer survivorship: improving care and quality of life.* 2003.
14. Oeffinger, K.C. and M.M. Hudson, *Long-term complications following childhood and adolescent cancer: foundations for providing risk-based health care for survivors.* CA: a cancer journal for clinicians, 2004. **54**(4): p. 208-236.
15. Oeffinger, K.C., *Longitudinal risk-based health care for adult survivors of childhood cancer.* Current problems in cancer, 2003. **27**: p. 143-167.
16. Salz, T., et al., *Are we ready to predict late effects? A systematic review of clinically useful prediction models.* European Journal of Cancer, 2015. **51**(6): p. 758-766.
17. Robison, L.L. and W. Demark-Wahnefried, *Cancer survivorship: focusing on future research opportunities.* Cancer Epidemiology, Biomarkers & Prevention, 2011. **20**(10): p. 1994-1995.
18. Oeffinger, K.C. and W.H.B. Wallace, *Barriers to follow-up care of survivors in the United States and the United Kingdom.* Pediatric blood & cancer, 2006. **46**(2): p. 135-142.

19. Dixon, S.B., et al., *Factors influencing risk-based care of the childhood cancer survivor in the 21st century.* CA: a cancer journal for clinicians, 2018. **68**(2): p. 133-152.
20. Guyon, I. and A. Elisseeff, *An introduction to variable and feature selection.* Journal of machine learning research, 2003. **3**(Mar): p. 1157-1182.
21. Shmueli, G. and O.R. Koppius, *Predictive analytics in information systems research.* MIS quarterly, 2011: p. 553-572.
22. Breiman, L., *Statistical modeling: The two cultures (with comments and a rejoinder by the author).* Statistical science, 2001. **16**(3): p. 199-231.
23. Shmueli, G., *To explain or to predict?* Statistical science, 2010. **25**(3): p. 289-310.
24. Varga, T.V., et al., *Association is not prediction: A landscape of confused reporting in diabetes–A systematic review.* diabetes research and clinical practice, 2020. **170**: p. 108497.
25. Steyerberg, E.W., M.J. Eijkemans, and J.D.F. Habbema, *Stepwise selection in small data sets: a simulation study of bias in logistic regression analysis.* Journal of clinical epidemiology, 1999. **52**(10): p. 935-942.
26. Sauerbrei, W., et al., *State of the art in selection of variables and functional forms in multivariable analysis—outstanding issues.* Diagnostic and prognostic research, 2020. **4**(1): p. 1-18.
27. Sun, G.-W., T.L. Shook, and G.L. Kay, *Inappropriate use of bivariable analysis to screen risk factors for use in multivariable analysis.* Journal of clinical epidemiology, 1996. **49**(8): p. 907-916.
28. Horan, M.R., et al., *A review of patient-reported outcome measures in childhood cancer.* Children, 2022. **9**(10): p. 1497.
29. Jacobsen, P.B. and H.S. Jim, *Consideration of quality of life in cancer survivorship research.* Cancer Epidemiology, Biomarkers & Prevention, 2011. **20**(10): p. 2035-2041.
30. Basch, E., et al., *Symptom monitoring with patient-reported outcomes during routine cancer treatment: a randomized controlled trial.* Journal of Clinical Oncology, 2016. **34**(6): p. 557.
31. Basch, E., et al., *Overall survival results of a trial assessing patient-reported outcomes for symptom monitoring during routine cancer treatment.* Jama, 2017. **318**(2): p. 197-198.
32. Denis, F., et al., *Two-year survival comparing web-based symptom monitoring vs routine surveillance following treatment for lung cancer.* Jama, 2019. **321**(3): p. 306-307.
33. Ayanian, J.Z. and P.B. Jacobsen, *Enhancing research on cancer survivors.* Journal of Clinical Oncology, 2006. **24**(32): p. 5149-5153.
34. Emery, J., et al., *Management of common clinical problems experienced by survivors of cancer.* The Lancet, 2022. **399**(10334): p. 1537-1550.
35. Landier, W., et al., *Development of risk-based guidelines for pediatric cancer survivors: The Children's Oncology Group long-term follow-up guidelines from the Children's Oncology Group Late Effects Committee and Nursing Discipline.* Journal of Clinical Oncology, 2004. **22**(24): p. 4979-4990.
36. Oeffinger, K.C., et al., *Chronic health conditions in adult survivors of childhood cancer.* New England Journal of Medicine, 2006. **355**(15): p. 1572-1582.
37. Phillips, S.M., et al., *Survivors of childhood cancer in the United States: prevalence and burden of morbidity.* Cancer Epidemiology, Biomarkers & Prevention, 2015. **24**(4): p. 653-663.

38. Hudson, M.M., et al., *Clinical ascertainment of health outcomes among adults treated for childhood cancer.* Jama, 2013. **309**(22): p. 2371-2381.
39. Hudson, M.M., et al., *Prospective medical assessment of adults surviving childhood cancer: study design, cohort characteristics, and feasibility of the St. Jude Lifetime Cohort study.* 2011. **56**(5): p. 825-836.
40. Robison, L.L., et al., *The Childhood Cancer Survivor Study: a National Cancer Institute–supported resource for outcome and intervention research.* 2009. **27**(14): p. 2308.
41. Robison, L.L., et al., *Study design and cohort characteristics of the Childhood Cancer Survivor Study: a multi-institutional collaborative project.* 2002. **38**(4): p. 229-239.
42. Sidey-Gibbons, J.A. and C.J. Sidey-Gibbons, *Machine learning in medicine: a practical introduction.* BMC medical research methodology, 2019. **19**: p. 1-18.
43. Rajkomar, A., J. Dean, and I. Kohane, *Machine learning in medicine.* New England Journal of Medicine, 2019. **380**(14): p. 1347-1358.
44. Beam, A.L. and I.S. Kohane, *Big data and machine learning in health care.* Jama, 2018. **319**(13): p. 1317-1318.
45. Kuhn, M. and K. Johnson, *Feature engineering and selection: A practical approach for predictive models*. 2019: Chapman and Hall/CRC.
46. Steyerberg, E.W., *Clinical prediction models*. 2009, New York: Springer Science+Business Media.
47. Nunez, E., E.W. Steyerberg, and J. Nunez, *Regression modeling strategies.* Revista Española de Cardiología (English Edition), 2011. **64**(6): p. 501-507.
48. Hastie, T., et al., *The elements of statistical learning: data mining, inference, and prediction*. Vol. 2. 2009: Springer.
49. Choi, Y., R. Park, and M. Seo, *Lasso on Categorical Data.* 2012.
50. Meier, L., S. Van De Geer, and P. Bühlmann, *The group lasso for logistic regression.* Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2008. **70**(1): p. 53-71.
51. Friedman, J., T. Hastie, and R. Tibshirani, *A note on the group lasso and a sparse group lasso.* arXiv preprint arXiv:1001.0736, 2010.
52. Simon, N., et al., *A sparse-group lasso.* Journal of Computational and Graphical Statistics, 2013. **22**(2): p. 231-245.
53. Zou, H. and T. Hastie, *Regularization and variable selection via the elastic net.* Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2005. **67**(2): p. 301-320.
54. Friedman, J., T. Hastie, and R. Tibshirani, *The elements of statistical learning*. Vol. 1. 2001: Springer series in statistics Springer, Berlin.
55. Newey, W.K. and D. McFadden, *Large sample estimation and hypothesis testing.* Handbook of econometrics, 1994. **4**: p. 2111-2245.
56. Zhao, P. and B. Yu, *On model selection consistency of Lasso.* The Journal of Machine Learning Research, 2006. **7**: p. 2541-2563.
57. Jia, J. and B. Yu, *On model selection consistency of the elastic net when $p \gg n$.* Statistica Sinica, 2010: p. 595-611.
58. Kass, R.E. and A.E. Raftery, *Bayes factors.* Journal of the american statistical association, 1995. **90**(430): p. 773-795.
59. Raftery, A.E., *Approximate Bayes factors and accounting for model uncertainty in generalised linear models.* Biometrika, 1996. **83**(2): p. 251-266.

60. Friedman, J., T. Hastie, and R. Tibshirani, *Regularization paths for generalized linear models via coordinate descent.* Journal of statistical software, 2010. **33**(1): p. 1.
61. Sauerbrei, W., et al., *On stability issues in deriving multivariable regression models.* Biometrical Journal, 2015. **57**(4): p. 531-555.
62. Heinze, G., C. Wallisch, and D. Dunkler, *Variable selection–a review and recommendations for the practicing statistician.* Biometrical journal, 2018. **60**(3): p. 431-449.
63. Kalousis, A., J. Prados, and M. Hilario, *Stability of feature selection algorithms: a study on high-dimensional spaces.* Knowledge and information systems, 2007. **12**(1): p. 95-116.
64. Meinshausen, N., *Relaxed lasso.* Computational Statistics & Data Analysis, 2007. **52**(1): p. 374-393.
65. Miller, K.D., et al., *Cancer treatment and survivorship statistics, 2022.* CA: a cancer journal for clinicians, 2022. **72**(5): p. 409-436.
66. Howlader, N., et al., *SEER cancer statistics review, 1975–2018.* National Cancer Institute, 2021.
67. Bhakta, N., et al., *The cumulative burden of surviving childhood cancer: an initial report from the St Jude Lifetime Cohort Study (SJLIFE).* 2017. **390**(10112): p. 2569-2582.
68. Robison, L.L. and M.M.J.N.R.C. Hudson, *Survivors of childhood and adolescent cancer: life-long risks and responsibilities.* 2014. **14**(1): p. 61-70.
69. Oeffinger, K.C., et al., *Chronic health conditions in adult survivors of childhood cancer.* 2006. **355**(15): p. 1572-1582.
70. Biddell, C.B., et al., *Developing personalized survivorship care pathways in the United States: Existing resources and remaining challenges.* Cancer, 2021. **127**(7): p. 997-1004.
71. Tralongo, P., M.S. McCabe, and A. Surbone, *Challenge for cancer survivorship: improving care through categorization by risk.* J Clin Oncol, 2017. **35**(30): p. 3516-7.
72. Nekhlyudov, L., et al., *Going beyond being lost in transition: a decade of progress in cancer survivorship.* Journal of Clinical Oncology, 2017. **35**(18): p. 1978.
73. Deshpande, P.R., et al., *Patient-reported outcomes: a new era in clinical research.* Perspectives in clinical research, 2011. **2**(4): p. 137.
74. Huang, I.-C., et al., *Association between the prevalence of symptoms and health-related quality of life in adult survivors of childhood cancer: a report from the St Jude Lifetime Cohort study.* 2013. **31**(33): p. 4242.
75. Gotay, C.C., et al., *The prognostic significance of patient-reported outcomes in cancer clinical trials.* 2008. **26**(8): p. 1355-1363.
76. Montazeri, A.J.H. and q.o.l. outcomes, *Quality of life data as prognostic indicators of survival in cancer patients: an overview of the literature from 1982 to 2008.* 2009. **7**(1): p. 102.
77. Sloan, J.A., et al., *Integrating patient-reported outcomes into cancer symptom management clinical trials supported by the National Cancer Institute–sponsored clinical trials networks.* 2007. **25**(32): p. 5070-5077.
78. Cantrell, M.A.J.J.o.P.O.N., *A narrative review summarizing the state of the evidence on the health-related quality of life among childhood cancer survivors.* 2011. **28**(2): p. 75-82.
79. Deshpande, P.R., et al., *Patient-reported outcomes: a new era in clinical research.* 2011. **2**(4): p. 137.

80. Robison, L.L., et al., *Study design and cohort characteristics of the Childhood Cancer Survivor Study: a multi-institutional collaborative project.* Medical and Pediatric Oncology: The Official Journal of SIOP—International Society of Pediatric Oncology (Societé Internationale d'Oncologie Pédiatrique, 2002. **38**(4): p. 229-239.
81. Hudson, M.M., et al., *Prospective medical assessment of adults surviving childhood cancer: study design, cohort characteristics, and feasibility of the St. Jude Lifetime Cohort study.* Pediatric blood & cancer, 2011. **56**(5): p. 825-836.
82. Brazier, J.E., et al., *Validating the SF-36 health survey questionnaire: new outcome measure for primary care.* 1992. **305**(6846): p. 160-164.
83. Ware, J.E., M. Kosinski, and S. Keller, *SF-36 physical and mental health summary scales: a user's manual*. 1994: Health Assessment Lab.
84. Ware, J., M. Kosinski, and S. Keller, *SF-36 physical and mental health summary scales.* A user's manual, 2001. **1994**.
85. Schulte, F., et al., *Development and Validation of Models to Predict Poor Health-Related Quality of Life Among Adult Survivors of Childhood Cancer.* JAMA network open, 2022. **5**(8): p. e2227225-e2227225.
86. Force, L.M., et al., *The global burden of childhood and adolescent cancer in 2017: an analysis of the Global Burden of Disease Study 2017.* The Lancet Oncology, 2019. **20**(9): p. 1211-1225.
87. Phillips, S.M., et al., *Survivors of childhood cancer in the United States: prevalence and burden of morbidity.* Cancer Epidemiol Biomarkers Prev, 2015. **24**(4): p. 653-63.
88. Bhakta, N., et al., *Cumulative burden of cardiovascular morbidity in paediatric, adolescent, and young adult survivors of Hodgkin's lymphoma: an analysis from the St Jude Lifetime Cohort Study.* The Lancet Oncology, 2016. **17**(9): p. 1325-1334.
89. Robison, L.L. and M.M. Hudson, *Survivors of childhood and adolescent cancer: life-long risks and responsibilities.* Nat Rev Cancer, 2014. **14**(1): p. 61-70.
90. Armstrong, G.T., et al., *Modifiable risk factors and major cardiac events among adult survivors of childhood cancer.* J Clin Oncol, 2013. **31**(29): p. 3673-80.
91. Darby, S.C., et al., *Risk of ischemic heart disease in women after radiotherapy for breast cancer.* N Engl J Med, 2013. **368**(11): p. 987-98.
92. Hudson, M.M., et al., *Clinical ascertainment of health outcomes among adults treated for childhood cancer.* JAMA, 2013. **309**(22): p. 2371-2381.
93. Mueller, S., et al., *Risk of first and recurrent stroke in childhood cancer survivors treated with cranial and cervical radiation therapy.* Int J Radiat Oncol Biol Phys, 2013. **86**(4): p. 643-8.
94. van der Pal, H.J., et al., *High risk of symptomatic cardiac events in childhood cancer survivors.* J Clin Oncol, 2012. **30**(13): p. 1429-37.
95. Ganz, P.A., et al., *Examining the influence of beta blockers and ACE inhibitors on the risk for breast cancer recurrence: results from the LACE cohort.* Breast Cancer Res Treat, 2011. **129**(2): p. 549-56.
96. Armstrong, G.T., M. Stovall, and L.L. Robison, *Long-term effects of radiation exposure among adult survivors of childhood cancer: results from the childhood cancer survivor study.* Radiat Res, 2010. **174**(6): p. 840-50.
97. Haugnes, H.S., et al., *Cardiovascular risk factors and morbidity in long-term survivors of testicular cancer: a 20-year follow-up study.* J Clin Oncol, 2010. **28**(30): p. 4649-57.

98. Meacham, L.R., et al., *Cardiovascular risk factors in adult survivors of pediatric cancer--a report from the childhood cancer survivor study.* Cancer Epidemiol Biomarkers Prev, 2010. **19**(1): p. 170-81.
99. Meadows, A.T., et al., *Second neoplasms in survivors of childhood cancer: findings from the Childhood Cancer Survivor Study cohort.* J Clin Oncol, 2009. **27**(14): p. 2356-62.
100. Mody, R., et al., *Twenty-five-year follow-up among survivors of childhood acute lymphoblastic leukemia: a report from the Childhood Cancer Survivor Study.* Blood, 2008. **111**(12): p. 5515-23.
101. Steffens, M., et al., *Endocrine and metabolic disorders in young adult survivors of childhood acute lymphoblastic leukaemia (ALL) or non-Hodgkin lymphoma (NHL).* Clin Endocrinol (Oxf), 2008. **69**(5): p. 819-27.
102. Turcotte, L.M., et al., *Temporal Trends in Treatment and Subsequent Neoplasm Risk Among 5-Year Survivors of Childhood Cancer, 1970-2015.* JAMA, 2017. **317**(8): p. 814-824.
103. Armstrong, G.T., et al., *Occurrence of multiple subsequent neoplasms in long-term survivors of childhood cancer: a report from the childhood cancer survivor study.* J Clin Oncol, 2011. **29**(22): p. 3056-64.
104. Bhatti, P., et al., *Risk of second primary thyroid cancer after radiotherapy for a childhood cancer in a large cohort study: an update from the childhood cancer survivor study.* Radiat Res, 2010. **174**(6): p. 741-52.
105. Bassal, M., et al., *Risk of selected subsequent carcinomas in survivors of childhood cancer: a report from the Childhood Cancer Survivor Study.* J Clin Oncol, 2006. **24**(3): p. 476-83.
106. Bhatia, S. and C. Sklar, *Second cancers in survivors of childhood cancer.* Nat Rev Cancer, 2002. **2**(2): p. 124-32.
107. Hudson, M.M., et al., *Long-term follow-up care for childhood, adolescent, and young adult cancer survivors.* Pediatrics, 2021. **148**(3).
108. Hudson, M.M., et al., *Approach for Classification and Severity Grading of Long-term and Late-Onset Health Events among Childhood Cancer Survivors in the St. Jude Lifetime Cohort.* Cancer Epidemiol Biomarkers Prev, 2017. **26**(5): p. 666-674.
109. Hudson, M.M., et al., *Approach for classification and severity grading of long-term and late-onset health events among childhood cancer survivors in the St. Jude Lifetime Cohort.* Cancer Epidemiology, Biomarkers & Prevention, 2017. **26**(5): p. 666-674.
110. Hastie, T., R. Tibshirani, and M. Wainwright, *Statistical learning with sparsity: the lasso and generalizations*. 2015: CRC press.
111. Obozinski, G., L. Jacob, and J.-P. Vert, *Group lasso with overlaps: the latent group lasso approach.* arXiv preprint arXiv:1110.0413, 2011.
112. Claggett, B., et al., *Comparison of Time-to-First Event and Recurrent-Event Methods in Randomized Clinical Trials.* Circulation, 2018. **138**(6): p. 570-577.
113. Kalbfleisch, J.D. and R.L. Prentice, *The statistical analysis of failure time data*. Vol. 360. 2011: John Wiley & Sons.
114. Fine, J.P. and R.J. Gray, *A proportional hazards model for the subdistribution of a competing risk.* Journal of the American statistical association, 1999. **94**(446): p. 496-509.

115. Pepe, M.S. and M. Mori, *Kaplan—meier, marginal or conditional probability curves in summarizing competing risks failure time data?* Statistics in medicine, 1993. **12**(8): p. 737-751.
116. Ghosh, D. and D.Y. Lin, *Marginal regression models for recurrent and terminal events.* Statistica Sinica, 2002: p. 663-688.
117. Rauh, S., *Survivorship Care for Cancer Patients: A Clinician's Handbook*. 2021: Springer Nature.
118. Gotay, C.C., et al., *The prognostic significance of patient-reported outcomes in cancer clinical trials.* Journal of clinical Oncology, 2008. **26**(8): p. 1355-1363.
119. Taylor, J. and R.J. Tibshirani, *Statistical learning and selective inference.* Proceedings of the National Academy of Sciences, 2015. **112**(25): p. 7629-7634.