

Evaluating *a priori* and data-driven weighting of the Healthy Eating Food Index-2019 for assessing diet quality and gastrointestinal and aerodigestive cancer risk in Canadian adults

by

Navreet Singh

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Science
in
Public Health Sciences

Waterloo, Ontario, Canada, 2025

© Navreet Singh 2025

Author's Declaration

This thesis consists of material all of which I authored or co-authored: see Statement of Contributions included in the thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Statement of Contributions

This thesis includes a manuscript that has been prepared for publication. Exceptions to sole authorship include:

Chapter 4: Navreet Singh¹, Abel Torres-Espin^{1,2}, Michael P. Wallace³, Lisa M. Bodnar⁴, Rachel A. Murphy^{5,6}, Marissa M. Shams-White^{7,8}, Sharon I. Kirkpatrick¹.

1. School of Public Health Sciences, University of Waterloo, Waterloo, ON, Canada
2. Department of Neurosurgery, University of California San Francisco, San Francisco, CA, USA
3. Statistics and Actuarial Sciences, University of Waterloo, Waterloo, ON, Canada
4. Department of Epidemiology, University of Pittsburgh School of Public Health, Pittsburgh, PA, USA
5. School of Population and Public Health, University of British Columbia, Vancouver, BC, Canada
6. Population Health Sciences, BC Cancer Research Institute, Vancouver, BC, Canada
7. Department of Population Science, American Cancer Society, Atlanta GA, USA
8. Division of Cancer Control and Population Sciences, National Cancer Institute, Bethesda, MD, USA

As lead author of **Chapter 4**, I was responsible for conceptualizing, analyzing and interpreting the data, as well as drafting the manuscript, with significant editorial contributions from my supervisor, Dr. Sharon Kirkpatrick.

Under Dr. Sharon Kirkpatrick's supervision, I also prepared the remaining chapters in this thesis, which were not written for publication.

Abstract

Background: Diet is a modifiable exposure implicated in gastrointestinal and aerodigestive cancers. Because foods are consumed in combination, diet quality indices are used to summarize overall dietary patterns. The Healthy Eating Food Index-2019 (HEFI-2019) measures adherence to Canada's Food Guide 2019, and its component scores are nearly equally weighted, reflecting the importance of all foods in a healthful dietary pattern. Its discriminatory capacity for measuring diet-disease associations, and the influence of the weighting schema of the index, remains uncertain.

Objective: To assess whether associations between diet quality and gastrointestinal and aerodigestive cancer risk differ among adults in Canada based on the *a priori* Healthy Eating Food Index-2019 (HEFI-2019) versus a novel modified version with components reweighted using a data-driven approach.

Methods: A prospective cohort analysis was conducted using the Canadian Community Health Survey 2004 Nutrition (CCHS 2004) linked with the Canadian Cancer Registry (CCR) through 2016. After exclusions, 10,530 adults were included, representing approximately 23.5 million Canadians. Diet was assessed using interviewer-administered 24-hour recalls. HEFI-2019 total scores were computed using standard weights and using data-driven weights derived from ridge-penalized Cox models in 10 iterations of 80/20 training–test splits with cross-validated penalty selection. Weighted Cox proportional hazards models, adjusted for age, sex, education, income, marital status, smoking status, body mass index, and alcohol consumption, estimated associations with incident gastrointestinal and aerodigestive cancers (ICD-9 140–149, 150–159, 160–161). Discrimination was assessed with Harrell's C-index.

Results: The data-driven approach altered component weights substantially (e.g., protein foods increased from 5 to 16.4; vegetables and fruits decreased from 20 to 3.73). No associations with cancer risk were observed for either the *a priori* (adjusted HR per unit increase 1.01; 95% CI: 0.99, 1.04) or reweighted HEFI-2019 scores (adjusted HR: 1.00; 95% CI: 0.98, 1.02). Model discrimination was similar (Harrell's C-index: 0.81 [95% CI: 0.77, 0.85] for *a priori*; 0.87 [95% CI: 0.80, 0.93] for reweighted).

Discussion: Neither the *a priori* nor reweighted HEFI-2019 was associated with gastrointestinal and aerodigestive cancer risk. Data-driven reweighting did not meaningfully improve associations or discriminatory capacity. These findings suggest challenges in using diet quality indices for complex diet-disease relationships and highlight the need for further research on index construction and application in cancer epidemiology.

Acknowledgements

I would first like to thank Dr. Sharon Kirkpatrick for her invaluable mentorship, support, guidance, and kindness over the past 5 years. There is a short list of mentors I can thank for having a transformative impact not only on my academic journey thus far, but for also shaping the way I see and participate in the world. Sharon, you are at the top of that list. I am so grateful that our paths crossed in the winter of 2020, and that you gave me the opportunity of taking my first step in research. You have helped me lay the foundation of my childhood dream of being a scientist by teaching me the importance of integrity and care in my work. You always supported my academic and career related pursuits however you could, and I hope to emulate the support you've shown to me as I take the next step forward. Thank you, Sharon.

To my committee members Drs. Abel Torres Espin and Michael Wallace, thank you for the support you have provided me throughout my MSc project and beyond. Your statistical and analytical expertise were invaluable, and I could not have completed this work without your support. Thank you, both, for the care and attention you've given to my project, even though my research area was far disconnected from yours. Abel, it has been a tremendous pleasure collaborating with you not only on my thesis work, but, also through my journey of learning more about machine learning and artificial intelligence. Michael, your ability to distill complex statistical concepts into easy to grasp ideas is something I have appreciated and admired since I began participating in the team's stats chats, thank you.

Thank you to the Public Health Nutrition team members. Silver and Lisa, it was amazing having you as team members as I undertook this journey. I couldn't have asked for better office mates and will forever cherish the memories we made together. I am so happy I got to learn from you both, and even more excited to see all the things you go onto accomplish. Alex, Alyssa, Joy, Lesley, and Martin, thank you for the comradery and your willingness to help me learn and grow as a researcher. I am truly grateful I had the opportunity to work with and learn from you all.

Thank you, Anastasia, Ava, and Jacob. I am so happy we got to traverse through graduate school together. I will always hold the memories of working on class assignments, trivia nights, and the random things we got up to when we were procrastinating, close to my heart. Our time spent together will remain a highlight of my MSc.

Thank you, Prabhnoor, the love of my life, for the support you've given me. Words can't capture what you mean to me, and I will always be indebted to you for putting up with my antics. There isn't a person in the world who understands me like you do. Thank you for standing by me through this journey, even though it kept us apart.

To my parents, brother, grandparents, extended family, and friends, thank you, for always encouraging me to pursue difficult things. I am so happy to have your love and support

in all that I do, even if you don't always understand what or why I am doing something. I wouldn't be who I am without you all. Thank you.

Dedication

I dedicate this thesis to my late grandmother, Ajit Kaur. Thank you for teaching me to cherish my education and pushing me to be a lifelong learner. I would not have made it as far as I have in my academic journey without the lessons you instilled in me throughout my childhood.

Table of Contents

Author's Declaration.....	ii
Statement of Contributions	iii
Abstract	iv
Acknowledgements	vi
Dedication.....	viii
List of Figures	xi
List of Tables	xii
List of Abbreviations.....	xiii
Chapter 1: Overview and Scope	1
Chapter 2: Literature Review	3
2.1 Diet quality and health.....	3
2.2 Public health nutrition surveillance and monitoring	4
2.3 Classifying Dietary Patterns.....	5
2.4 Dietary indices	6
2.5 Considerations for using diet quality indices in epidemiological research	7
2.6 Index weighting and association with health outcomes	8
2.7 Dietary risk factors for gastrointestinal malignancies.....	9
2.8 Summary.....	11
Chapter 3: Study Rationale, Objectives and Hypothesis	12
Chapter 4: Evaluating <i>a priori</i> and data-driven weighting of the Healthy Eating Food Index-2019 for assessing diet quality and gastrointestinal and aerodigestive cancer risk in Canadian adults.....	13
4.1 Abstract	14
4.2 Introduction	15
4.3 Methods.....	18
4.3.1 Data	18
4.3.2 HEFI-2019.....	19
4.3.3 Event and time-to-event	23
4.3.4 Covariates	23
4.3.5 Statistical analyses.....	25
4.4 Results.....	27
4.5 Discussion	30
4.6 References for Chapter 4	38
Chapter 6: General Discussion.....	45

6.1 Summary of key findings.....	45
6.2 Strengths	46
6.3 Study limitations	46
6.4 Implications for future research.....	47
6.5 Conclusion.....	49
References	50
Appendix A: Supplementary materials for Chapter 4	55

List of Figures

Figure 4.1. Directed acyclic graph illustrating the hypothesized relationships between diet quality and gastrointestinal and aerodigestive cancers, including covariates, effect modifiers, and unmeasured factors.	24
--	----

List of Tables

Table 4.2. Baseline characteristics of participants by cancer case status, Canadian Community Health Survey–Nutrition 2004 (CCHS 2004)–Canadian Cancer Registry linked cohort (N=23,545,000	28
Table 4.3. Standard and Data-Driven Weights for HEFI-2019 Components Used to Assess Diet Quality in Relation to Gastrointestinal and Aerodigestive Cancer Risk	29
Table 4.4. Association of Standard and Data-Driven Weighted HEFI-2019 Scores with Gastrointestinal and Aerodigestive Cancer Risk in the CCHS 2004–CCR Linked Cohort (N=2,354,400)	30

List of Abbreviations

HEFI-2019	Healthy Eating Food Index-2019
HEI	Healthy Eating Index
CCHS 2004	Canadian Community Health Survey 2004 Nutrition
CCR	Canadian Cancer Registry
NHANES	National Health and Nutrition Examination Survey
HLI	Healthy Lifestyle Index
EPIC	European prospective investigation into cancer and nutrition

ਪੇਥੀ ਪੜਿ ਪੜਿ ਜਗ ਮੁਆ, ਪੰਡਿਤੁ ਭਇਆ ਨ ਕੋਇ।

ਢਾਈ ਅਖਰ ਪ੍ਰੇਮ ਕਾ, ਪੜੈ ਸੇ ਪੰਡਿਤੁ ਹੋਇ॥

“The world has perished reading countless books,
yet none became truly wise.
He alone is learned who understands love.”

— Kabir

Chapter 1: Overview and Scope

In Canada, the overall societal cost of cancer was projected to be \$37.7 billion in 2024, with greater than 240,000 new cancer cases diagnosed annually (1). Evidence from meta-analyses and prospective cohort studies consistently links high intakes of red and processed meats with increased risks for cancers of the esophagus, stomach, pancreas, and colorectum, while diets rich in vegetables, fruits, whole grains, and dietary fibre are protective (2–12). Although it is essential to understand the relationship between specific foods and cancer risk, nutrition research increasingly emphasizes that foods are consumed in combination, forming complex dietary patterns rather than isolated exposures (13).

Dietary patterns capture the amounts, proportions, variety, and combinations of foods, beverages, and nutrients regularly consumed, as well as their frequency of intake (13). Understanding diet at this level moves beyond single-food approaches and attempts to account for potential cumulative or synergistic effects of multiple dietary exposures (13). *A priori* indices are commonly used to measure alignment with nutritional guidance and characterize dietary patterns. In epidemiological research, such indices are used to describe dietary exposure and further understand the relationships between diet and disease (14,15). These indices specify dietary components, apply cutoffs or ratios for scoring, and aggregate the results into a total score that reflects adherence to guidance (14,15). Examples of such indices include the Healthy Eating Index (HEI) and, in Canada, the Healthy Eating Food Index-2019 (HEFI-2019) (16,17).

The HEFI-2019 consists of ten components, including five food-based, four nutrient-based, and one beverage component, each scored as adequacy (higher intakes yield higher scores) or moderation (lower intakes yield higher scores) (17). Total scores are calculated as a weighted sum of components, with vegetables and fruits assigned greater weight in line with Canada's Food Guide recommendations (17). Like other indices, the HEFI-2019 is designed for population monitoring and public health surveillance but can be applied in epidemiological studies linking diet quality with disease outcomes (18).

A key methodological issue, however, is how components of diet quality indices are weighted. Historically, most indices, including the HEFI-2019, have applied equal or near-equal weights, reflecting the assumption that all components contribute similarly to overall diet

quality (14,15). Yet emerging evidence suggests this approach may dilute associations with health outcomes by obscuring the influence of components more strongly related to disease risk (19–21). For example, studies have demonstrated that tailored or empirically derived weights can improve predictive capacity. In the European Prospective Investigation into Cancer and Nutrition (EPIC) study, reweighting components of a Healthy Lifestyle Index strengthened associations with type 2 diabetes incidence (20). Similarly, an empirically weighted Healthy Eating Index showed stronger associations with mortality risk compared to the standard index in a large cohort of postmenopausal women (21). These findings suggest that reweighting components according to their empirical contributions to disease risk may enhance discriminatory capacity.

Although diet quality indices are widely used in international research, evidence linking these measures with cancer outcomes remains inconsistent. The ability of diet quality indices to capture exposures most relevant to disease may be influenced by differences in design, including the selection and weighting of components. At the same time, the multifactorial nature of cancer and the complexity of dietary patterns underscore the challenges of detecting clear associations. These considerations underscore the importance of evaluating how diet quality indices perform in various populations and contexts, as well as whether modifications to their construction can enhance their utility in understanding disease risk.

This thesis examines whether associations between metrics of diet quality and the risk of gastrointestinal and aerodigestive cancers vary when the HEFI-2019 is applied using its standard *a priori* weights compared to empirically derived weights. Chapter 2 outlines the broader context for this work, followed by Chapter 3, which details the study rationale, objectives, and hypotheses. Chapter 4 presents the core manuscript prepared for submission to *Public Health Nutrition*, and Chapter 5 concludes with a general discussion of the findings, their strengths and limitations, and their implications for research and practice.

Chapter 2: Literature Review

2.1 Diet quality and health

Nutrition and health are intrinsically linked. Poor diet quality is a leading risk factor for chronic diseases worldwide (22,23). Globally in 2013, dietary risks accounted for 11.3 million deaths and 241.4 million disability-adjusted life-years (22). Poor diet quality has been linked to an increased risk of type 2 diabetes mellitus, cardiovascular disease, metabolic syndrome, and cancer (24). A systematic review and meta-analysis examining the relationship between diet quality and cardiovascular disease found a lower incidence of disease and mortality risk among individuals with high diet quality compared to those with low diet quality (25). In addition, a systematic review and meta-analysis found evidence of decreased risk of colon and breast cancer among individuals with a healthy dietary pattern (characterized by consumption of fruits, vegetables, legumes, whole grains, and fibre) as compared to those with unhealthy dietary patterns (characterized by consumption of red and processed meat, sugary drinks and salty snacks, starchy foods, and refined carbohydrates) (26). In Canada, failing to meet recommendations related to vegetable, fruit, whole grain, milk, and nut and seed consumption was estimated to result in an economic burden of \$13.8 billion in direct healthcare costs and indirect costs (27).

Modifying diet has been identified as an approach for reducing the burden of chronic diseases and promoting population health (28). Efforts to modify dietary patterns in a population typically involve promoting the consumption of specific foods and nutrients, while restricting or eliminating the intake of others (28). However, focusing on particular foods or nutrients can be viewed as a reductionist approach to evaluating the impact of diet on health, as individual-level diets are composed of many different foods consumed together (29). To capture the complexity of dietary patterns — defined as the foods, food groups, and nutrients, including their combinations, frequencies, and quantities, consumed by an individual or population (29) — governments globally have released food-based dietary recommendations (28, 30, 31). Food-based dietary guidelines aim to provide recommendations on healthy dietary patterns and are developed to be context-specific and well-rooted in the best available evidence (28). Canada's Food Guide 2019 is an example of food-based dietary guidance. Unlike

the prior iteration of Canada's Food Guide, the 2019 version does not include recommendations to consume specific numbers of servings of different types of foods in relation to individual energy needs. Instead, through a visual plate, Canada's Food Guide 2019 promotes consuming a variety of fruits and vegetables, whole grains, and protein foods, with a specific emphasis on plant-based protein foods (32). The Food Guide suggests that fruits and vegetables should comprise half of a meal, while grains and protein foods should each account for a quarter of a typical meal. The Food Guide also suggests consuming water and limiting intake of highly processed foods, such as sugar-sweetened beverages and processed meats (32).

2.2 Public health nutrition surveillance and monitoring

In addition to providing recommendations, a crucial component of promoting healthful dietary patterns includes surveillance to monitor the dietary intake of populations (16). Surveillance can serve to understand diet quality, including potential differences across a population and over time, identify at-risk groups, monitor health promotion efforts (e.g., Canada's Food Guide), and enable the analysis of risk factors for diseases (16). A common and practical approach to food and nutrition surveillance is the use of large-scale surveys that capture information regarding food and nutrition, health status, and other health-related variables (16). The U.S. National Health and Nutrition Examination Survey (NHANES) is an example of a large-scale survey (33). Administered continuously since 1999, the survey captures information regarding the prevalence of and risk factors for major diseases, assesses dietary intake, serves as the basis for national standards (e.g., for measurements such as height, weight, and blood pressure), and aims to enable the identification of undiagnosed conditions like diabetes (33). Similarly, in Canada, the Canadian Community Health Survey – Nutrition captures information on dietary intake. However, nutrition-focused versions of the Canadian Community Health Survey have been administered only twice, first in 2004 and more recently in 2015 (34,35).

Both NHANES and the Canadian Community Health Survey capture information on dietary intake using 24-hour recall methodology (33,34). In brief, 24-hour dietary recall methodology consists of a respondent recalling information regarding the foods and beverages

consumed on the day prior from midnight to midnight and can be self- or interviewer-administered (36). When linked with food composition databases that provide information on the nutrient composition of common foods and beverages, nutrient intake can be estimated (36). Such information can provide valuable insights into the intake of nutrients of concern, such as calcium, vitamin D, fibre, iron, and potassium. This information can be further leveraged to examine the association between particular nutrients or foods like vegetables and fruits, and specific health outcomes, or sociodemographic variables (36). For instance, Cohen and colleagues (2006) assessed the association between sodium intake and cardiovascular disease and all-cause mortality by leveraging NHANES data (37). However, when specific nutrients are evaluated in isolation, it is not possible to ascertain overall dietary patterns.

2.3 Classifying Dietary Patterns

Two approaches to describing dietary patterns using dietary intake data include *a posteriori* and *a priori* methods (38). The *a posteriori* approach leverages statistical methods, such as cluster analysis, to reveal common patterns of food consumption within a population (38,39). Data-driven methods do not rely on a definition of what constitutes healthy or unhealthy dietary patterns, and the derived dietary patterns are inherently multidimensional (14,39). However, identified patterns may not be related to specific health outcomes, and the researcher may need to make many subjective decisions (14,39). Specifically, depending on the statistical method, researchers may need to decide how items are grouped, how input variables are considered, what constitutes a meaningful association between input variables, and how many dietary patterns are present in the data (14,15). Identified patterns may not be reproducible across different populations, limiting the generalizability of a particular data-driven dietary pattern to other populations (38).

In contrast to *a posteriori* approaches, *a priori* approaches rely on current nutritional knowledge and are constructed based on defined dietary components identified as relevant to health promotion and a healthful dietary pattern (38). The *a priori* approach to investigating dietary patterns manifests as dietary indices (38,39). Some examples of dietary indices include the Diet Quality Index (40), the Healthy Eating Index (HEI) (41), and the Mediterranean Diet Score (42). Dietary intake data can be leveraged to compute an index score that can provide the

researcher with insights into how well an individual or population's diet aligns with the predefined dietary standard underlying the index's construction. Advantages to *a priori* approaches include higher interpretability, less computational demand, and higher reproducibility and comparability, compared to *a posteriori* or data-driven approaches (14,15). However, the reliance of dietary indices on existing nutrition-health knowledge presents a critical disadvantage, given that knowledge on what constitutes a healthful diet is incomplete (14,15). Furthermore, dietary indices should only be applied in scenarios for which they have been intended to be used (14,15). For instance, it may not be appropriate to use a dietary index developed for adults in a child population (14).

2.4 Dietary indices

Dietary indices are often explicitly leveraged in public health monitoring and surveillance applications, as they can be designed to measure adherence to nutritional guidelines. The HEI is an example of a dietary index intended to assess diet quality with respect to the Dietary Guidelines for Americans (43). Higher component scores reflect greater adherence to the DGA for that respective component, and higher aggregate scores reflect greater overall dietary quality (43).

The HEFI-2019 can be seen as the Canadian counterpart to the HEI (17). Similar in philosophy to the HEI, HEFI-2019 captures adherence to Canada's Food Guide 2019 healthy food choices recommendations and can be used as a proxy for diet quality to the extent that the guidance reflects a healthy diet (17). Seven HEFI-2019 components (vegetables and fruits, whole-grain foods, grain foods ratio, protein foods, plant-based protein foods, beverages, and fatty acid ratio) are considered adequacy-based, while three components (saturated fats, free sugars, and sodium) are moderation-based. Individuals can receive 5, 10, or 20 points depending on the component, for an aggregate total of 80 points (17). Similar to the HEI, higher component scores illustrate greater alignment to dietary guidance related to that component, and higher aggregate scores represent greater adherence to Canada's Food Guide 2019 healthy food choices recommendations overall (17).

The components within the HEFI-2019 are weighted. That is, specific components contribute more or less to the total score (17,41). For example, the vegetables and fruits

component of the HEFI-2019 is allocated a maximum of 20 points to align with Canada's Food Guide 2019, in which it is recommended that half of an individual's meal is composed of vegetables and fruits, and the scientific evidence that stresses the importance of vegetable and fruit consumption in a high-quality dietary pattern (17). Aside from the vegetables and fruits component, all other components of the HEFI-2019 are equally weighted (17).

2.5 Considerations for using diet quality indices in epidemiological research

Neither the HEI nor HEFI-2019 are explicitly intended to be used to assess diet quality-health associations, with the index score representing an independent variable and some health outcome representing a dependent variable. However, index scores are often used to assess diet-health relationships. Although diet quality indices rely on the best available evidence on the relationship between diet and health and disease prevention, this evidence base is often insufficient, and there is limited consensus on what constitutes a healthy dietary pattern with respect to minimizing risk for health conditions (14). Furthermore, since diet quality indices are based on dietary guidance tailored for the general population, their construction does not focus on reducing risk for a specific condition, such as cancer or cardiovascular disease (14). Accordingly, what constitutes a healthful dietary pattern for the general population (as captured by the HEFI-2019, for instance) may not necessarily be applicable in reducing risk for a health condition.

The use of the overall HEI or HEFI-2019 score as an independent variable does not provide insights into the food components that may individually be associated with the outcome of interest (44). This can be problematic since individuals with similar aggregate scores may have vastly different dietary patterns. For example, focusing on the HEFI-2019, two individuals could have different scores for the vegetables and fruits and the beverages component, with the same score for the remaining components, and still share the same aggregate score. Essentially, middling scores may be due to high scores for some components and lower scores on others, or middling scores on most components. It is thus recommended that, in addition to aggregate scores, component scores be modelled so that particular component associations to some outcome can be ascertained (44). However, analysis of

individual components alone may also be problematic, as it relinquishes the ability of the indices to account for potentially synergistic effects between various dietary components (44).

2.6 Index weighting and association with health outcomes

Particular dietary components may be more relevant to the risk of a disease outcome than others. A handful of studies have examined the appropriateness of equal weighting approaches of the components within diet quality indices when examining the relation between diet quality and a health outcome. Leveraging data from the Nulliparous Pregnancy Outcomes Study: Monitoring Mothers-to-Be (nuMoM2b), a prospective pregnancy cohort in the United States, Peterson and colleagues (2022) demonstrate that equal weighting of the HEI-2010 components may be inappropriate when assessing the relationship between diet quality and pregnancy-related outcomes like small-for-gestational-age birth, preterm birth, preeclampsia, and gestational diabetes (19). Their investigations revealed that components like greens and beans and total vegetables were more strongly associated with fewer cases of small-for-gestational-age birth, preterm birth, and preeclampsia than other components(19). These findings illustrate that a differential weighting scheme for the HEI may be appropriate when examining the relationship between diet quality and pregnancy outcomes (19).

Parker et al. (2024) found that data-driven modified-weight HEI-2015 scores were associated with a 23% to 38% reduction in all-cause mortality risk when comparing individuals in the highest HEI-2015 score quintile to those in the lowest quintile (45). Interestingly, they found no association between standard HEI-2015 scores and the risk of all-cause mortality or cardiovascular disease mortality (45). By employing LASSO regression methodology, the study reweighted HEI-2015 scores such that component scores were adjusted from a maximum of 5 or 10 points to new maximum component scores depending on the relevance of that component to the relationship between intake and all-cause mortality (45). For example, maximum scores for moderation components, such as added sugars and sodium, were reduced from 10 points each to 0 and 2.6 for women and 0.2 and 6.6 for men, respectively (45). Other components, like total fruits, were adjusted from a previous maximum of 5 points to a maximum of 17.2 points for women and 16.2 points for men (45). By employing a data-driven methodology to reweight specific dietary components, this investigation reveals that equal α

priori component weighting may not be appropriate when examining the relationship between diet quality and all-cause mortality (45). Investigations by Petersen et al. (2022), and Parker et al. (2024), reveal that what is understood to be a healthful dietary pattern for the general population may not hold when examining the relationship between diet quality and health outcomes (45).

Findings on the appropriateness of equal weighting of components within *a priori* indices to assess the relationship between certain individual-level health factors or behaviours and a particular health outcome or condition are not limited to diet quality. Leveraging data from the European Prospective Investigation into Cancer and Nutrition (EPIC) study, Viallon and colleagues (2024) revealed that the unweighted healthy lifestyle index (HLI) was inappropriate when leveraging the index to assess the association between lifestyle components and health-related outcomes (20). The HLI is an *a priori* health index that attempts to capture various health-related factors (20). Specifically, it takes into consideration an individual's smoking status, alcohol consumption, diet, physical activity levels, and adiposity (20). Each component is scored in a binary manner, with a maximum overall score of 5. Viallon and colleagues (2024) reweighted each component based on Cox regression parameters for the main effects and the interaction terms of each HLI component in relation to a health outcome (20). Their findings revealed that the hazard ratios for a one standard deviation increase in the standard and type 2 diabetes mellitus specific HLI were 0.66 (95% CI= 0.64, 0.67) and 0.43 (95% CI= 0.42, 0.44), respectively (20). In the type 2 diabetes-specific HLI, factors such as adiposity were weighted more heavily, while other factors, including smoking and diet, were weighted less (20). Collectively, these studies suggest that a more nuanced approach to weighting is required when using indices to assess relationships between aggregate index scores and health outcomes.

2.7 Dietary risk factors for gastrointestinal malignancies

Gastrointestinal malignancies include cancers of the gastrointestinal (i.e., esophagus, stomach, pancreas, gallbladder, colon, rectum) and aerodigestive (i.e., mouth, pharynx, larynx) tracts. Collectively, gastrointestinal malignancies represent close to one in six of all cancer cases in Canada (46). In 2024, 43,700 new gastrointestinal cancer cases were projected, with nearly 90% of all cancer diagnoses occurring in individuals over the age of 50 (46). The strength of

evidence highlighting the relationship between specific foods or dietary patterns and gastrointestinal cancers varies depending on the anatomical site; however, there is considerable overlap of risk and protective factors across sites. High consumption of red and processed meats has repeatedly been linked to elevated risks of several cancers, including esophageal squamous cell carcinoma, stomach, pancreatic, and prostate cancers in multiple meta-analyses and cohort studies (2–4). A 2024 systematic review and meta-analysis of 95 cohort studies including 5,794,219 participants identified elevated risks for esophageal, pancreatic, liver, colon, and colorectal cancers, among those in the highest versus lowest quantiles of red and processed meat consumption (5). Further, the International Agency for Research on Cancer (IARC) has classified processed meat as a Group 1 carcinogen, with each 50g/day increase in consumption associated with an 18% increase in colorectal cancer risk (2). Consumption of salt-preserved, smoked, pickled, or grilled foods, which include nitrosamines and polycyclic aromatic hydrocarbons, is associated with increased risk of gastric cancer (6,47).

In contrast, diets rich in fruits, non-starchy vegetables, and fibre have been found to be protective against cancers of the oral cavity, pharynx, larynx, esophagus, stomach, pancreas, gallbladder, and small intestine (6–12). Meta-analyses of case-control studies indicate ~40-50% lower risk of oral cavity and pharyngeal cancers for individuals at the highest fruit and vegetable intakes, while prospective cohort studies report smaller but consistent risk reductions (~20-30%), with similar but more varied findings for laryngeal cancer (48–50). For colorectal cancer, dietary fibre, particularly from whole grains, has been found to reduce colorectal cancer risk by 17% (95% CI: 11-22%) for each 90g/day serving (51). Dairy product intake may also be protective against colorectal cancer, with ~400g/day of total dairy lowering risk by 13% (95% CI: 10-17%) (52). At the dietary pattern level, adherence to a Mediterranean dietary pattern rich in fruits, vegetables, legumes, whole grains, nuts, and olive oil, and low in salt and processed meat, demonstrates strong protective associations for oral cavity, pharynx, larynx, gastric, and colorectal cancers (8,12). Ultimately, foods like fruits, vegetables, and whole grains may be protective factors against gastrointestinal malignancies, while the consumption of other foods, such as red and processed meats, may be a risk factor for these conditions.

2.8 Summary

Substantial evidence suggests that dietary patterns play a significant role in influencing the risk of gastrointestinal and aerodigestive cancers. High consumption of red and processed meats, as well as salt-preserved or energy-dense foods, is associated with an increased risk of cancer, while diets rich in fruits, vegetables, whole grains, and dairy appear to be protective across multiple sites. Moreover, adherence to dietary patterns such as the Mediterranean diet confers additional reductions in cancer risk. Despite this, uncertainty remains regarding how best to capture and quantify diet-cancer relationships, particularly in terms of diet quality indices. This gap provides the foundation for the present thesis, which seeks to evaluate whether standard *a priori* versus data-driven weighting of a dietary index influences associations with gastrointestinal and aerodigestive cancer risk.

Chapter 3: Study Rationale, Objectives and Hypothesis

The objective of this thesis was to investigate whether the associations between diet quality and the risk of gastrointestinal and aerodigestive cancers among Canadian adults vary depending on how diet quality is operationalized. Specifically, the thesis examines the HEFI-2019, which was developed to measure adherence to Canada's Food Guide 2019. While the HEFI-2019 provides a context-specific measure of diet quality for the Canadian population, its current weighting scheme reflects general dietary guidance rather than evidence tailored to disease prevention. As such, questions remain about whether the *a priori* weighting approach embedded in the HEFI-2019 is optimal for detecting relationships between dietary patterns and cancer risk.

To address this gap, this thesis applies both the standard HEFI-2019 and a modified version in which component weights are re-estimated using a data-driven penalized regression approach. This comparative approach enables the assessment of whether reweighting dietary components based on their statistical relevance to cancer outcomes enhances the index's capacity to capture meaningful associations.

It is hypothesized that the modified HEFI-2019 with reweighted components will demonstrate stronger associations with cancer incidence and greater discriminatory performance than the standard *a priori* index. By testing this hypothesis, this thesis seeks to provide insight into how different approaches to weighting diet quality indices may influence observed associations between diet and cancer risk. The findings may help clarify methodological considerations for using dietary indices in nutritional epidemiology and inform future research and public health applications.

Chapter 4: Evaluating *a priori* and data-driven weighting of the Healthy Eating Food Index-2019 for assessing diet quality and gastrointestinal and aerodigestive cancer risk in Canadian adults

In preparation for submission to *Public Health Nutrition (PHN)*.

4.1 Abstract

Objective: To assess whether associations between diet quality and gastrointestinal and aerodigestive cancer risk differ among adults in Canada based on the *a priori* Healthy Eating Food Index-2019 (HEFI-2019) versus a novel modified version with components reweighted using a data-driven approach.

Design: Diet quality was assessed using HEFI-2019 total scores (0-80 points), and with components reweighted via ridge regression in 10 train-test splits. Associations with cancer risk were examined using weighted Cox proportional hazards models adjusted for age, sex, educational attainment, income, marital status, smoking status, alcohol consumption, and body mass index.

Setting: Canada.

Participants: 10,530 adults (≥ 18 y) representing 23,545,000 weighted participants of whom 383,000 developed incident gastrointestinal and aerodigestive cancer.

Results: The data-driven approach altered component weights substantially (e.g., protein foods increased from 5 to 16.4; vegetables and fruits decreased from 20 to 3.73). No associations with cancer risk were observed for either the *a priori* (adjusted HR per unit increase: 1.01; 95% CI: 0.99, 1.04) or reweighted HEFI-2019 scores (adjusted HR: 1.00; 95% CI: 0.98, 1.02). Model discrimination was similar (Harrell's C-index: 0.81 [95% CI: 0.77, 0.85] for *a priori*; 0.87 [95% CI: 0.80, 0.93] for reweighted).

Conclusions: Neither the *a priori* nor reweighted HEFI-2019 total score was associated with gastrointestinal and aerodigestive cancer risk. Data-driven reweighting did not meaningfully improve associations or discriminatory capacity. These findings suggest challenges in using diet quality indices for complex diet-disease relationships and highlight the need for further research on index construction and application in cancer epidemiology.

4.2 Introduction

The overall societal cost of cancer in Canada, including healthcare costs, patient-out-of-pocket costs, and lost productivity, was projected to be \$37.7 billion in 2024 (1). Collectively, gastrointestinal malignancies represent close to one in six of all cancer cases nationally (2). Gastrointestinal malignancies include cancers of the gastrointestinal (i.e., esophagus, stomach, pancreas, gallbladder, colon, rectum) and aerodigestive (i.e., mouth, pharynx, larynx) tracts. In 2024, 43,700 new gastrointestinal cancer cases were projected, with nearly 90% of all cancer diagnoses occurring in individuals over the age of 50 (2).

Diet is believed to play a substantial role in the development of gastrointestinal and aerodigestive cancers. Red and processed meat have consistently been identified as key dietary risk factors, with evidence linking higher consumption to esophageal squamous cell carcinoma, stomach, pancreatic, and prostate cancers in multiple meta-analyses and cohort studies (3–5). A 2024 meta-analysis of 95 cohort studies (n=5,794,219) found elevated risks for esophageal, pancreatic, liver, colon, rectal, and colorectal cancers, among those in the highest versus lowest quantiles of red and processed meat consumption (6). Diets high in salt-preserved, pickled, or smoked foods are associated with elevated risks for gastric cancer (7,8). Conversely, diets rich in fruits, non-starchy vegetables, and fibre have been found to be protective against cancers of the oral cavity, pharynx, larynx, esophagus, stomach, pancreas, gallbladder, and small intestine (8–14). Dietary fibre, particularly from whole grains, is strongly associated with lower colorectal cancer risk, and dairy products and calcium may also protect against colorectal cancer (8,15).

Although it is helpful to understand the relationship between specific foods or food groups and cancer risk, individuals do not consume foods in isolation. Nutrition research has increasingly emphasized dietary patterns, defined as the amounts, proportions, diversity, or combinations of foods, beverages, and nutrients regularly consumed, along with their frequency of intake (16). Examining overall dietary patterns offers a way to capture the potential synergistic and cumulative effects of multiple dietary exposures on cancer risk and moves beyond single-exposure approaches that may overlook complex interactions (16). Two approaches to describing dietary patterns using dietary intake data include *a posteriori* and *a priori* methods (17). The *a posteriori* approach leverages statistical methods, such as principal

component analysis and cluster analysis, to reveal common patterns of food consumption within a population (17,18). These data-driven methods do not rely on preconceived definitions of healthy or unhealthy dietary patterns, and the derived patterns are inherently multidimensional (18,19). In contrast, *a priori* approaches rely on current nutrition knowledge and are constructed based on predefined dietary components that are identified as relevant to a healthful dietary pattern (17).

To assess dietary patterns on a population-level, researchers often rely on *a priori* diet quality indices due to their standardization and alignment with public health recommendations (19,20). Common examples include the Healthy Eating Index (HEI) and the Healthy Eating Food Index-2019 (HEFI-2019) (21,22). These indices share a common structure: a set of predefined components (e.g., vegetables and fruits, whole grains, free sugars, sodium) that are each scored using cut-points or ratios and then combined into a total score as a weighted sum. Components typically fall into “adequacy” (higher intake earns more points) and “moderation” (lower intake earns more points) categories. Historically, components have been weighted equally (or mostly equally) to reflect the principle that dietary guidance should be taken as a whole, based on the assumption that all components contribute similarly to overall diet quality (19,20). For instance, HEFI-2019 weights all components equally except for the vegetable and fruits component, which is allocated a maximum of 20 points to align with Canada’s Food Guide 2019 recommendation that half an individual’s typical meal should be composed of vegetables and fruits (22). Although designed to capture adherence to dietary guidance, diet quality indices are widely used in epidemiological research to examine associations with health outcomes, including cancers (23). However, given that diet quality indices focus on broad dietary guidance rather than disease-specific evidence, component selection and weighting may not fully capture the dietary factors most relevant to cancer risk, potentially diluting or obscuring associations between overall diet quality and disease risk.

Growing evidence suggests that equal weighting of components in diet quality indices and other health-related indices may mask important associations with health outcomes. In the Nulliparous Pregnancy Outcomes Study: Monitoring Mothers-to-Be (nuMoM2b), a prospective pregnancy cohort in the United States, Peterson and colleagues (2022) found that HEI-2010

components like greens and beans and total vegetables were more strongly associated with reduced risk of adverse pregnancy outcomes than other components, indicating that uniform weighting may be inappropriate (24). Viallon et al. (2024), using data from the European Prospective Investigation into Cancer and Nutrition (EPIC) study, reweighted components of the Healthy Lifestyle Index (HLI), which captures the combined effects of smoking, alcohol consumption, diet, physical activity, and adiposity on health (25). Stronger associations with type 2 diabetes were found when adiposity received greater emphasis than in the standard weighting (25). The hazard ratios for a one standard deviation increase in the standard and type 2 diabetes mellitus specific HLI were 0.66 (95% CI= 0.64, 0.67) and 0.43 (95% CI= 0.42, 0.44), respectively, based on the modified and standard weighting (25). Prentice et al. (2022) developed an empirically weighted Healthy Eating Index (E-HEI) using Cox regression coefficients for calibrated dietary variables to improve prediction of mortality in a cohort of postmenopausal women (26). Compared to the lowest quintile, women in the highest E-HEI quintile had a 26% reduction in cancer mortality and a 28% reduction in all-cause mortality (26). In contrast, the standard HEI-2010 showed weaker and less consistent associations (26). Component weights in the E-HEI were optimized based on their empirical contribution to mortality risk, resulting in down-weighting of components such as seafood and plant protein, and up-weighting of components like whole grains and whole fruit (26). Collectively, these findings support a more tailored approach to how indices representative of complex constructs such as overall diet quality are operationalized when evaluating disease risk.

The HEFI-2019 was developed to measure adherence to the healthy food choices recommendations within Canada's Food Guide 2019. Although its application in epidemiological research is emerging (27,28), it offers a context-specific measure of diet quality for the Canadian population. However, its weighting scheme is based on general dietary guidance rather than disease-specific evidence, and it remains unclear whether this weighting approach is appropriate for assessing diet-cancer risk. The objective of this study was thus to assess whether associations between diet quality and gastrointestinal and aerodigestive cancer risk differ among adults in Canada based on the *a priori* HEFI-2019 versus a novel modified version with components reweighted using a data-driven approach. We hypothesize that the modified

HEFI-2019 with reweighted components will demonstrate stronger associations with cancer risk and offer greater discriminatory capacity, as compared to the *a priori* HEFI-2019 among Canadian adults.

4.3 Methods

4.3.1 Data

Data were drawn from the linked dataset formed by the Canadian Community Health Survey – Nutrition 2004 (CCHS 2004) and the Canadian Cancer Registry (CCR). While the Canadian Community Health Survey (CCHS) is administered annually, detailed dietary intake data were collected only in 2004 and 2015. Data from the 2004 wave are used in the current analysis to allow for time between dietary intake measurement and disease onset, and because the linkage is only available up to the diagnosis year 2016 (29).

CCHS employs a multistage stratified cluster sampling design that uses individual dwellings as the sampling unit (30). The stratified design ensures that the sample is representative with respect to age, sex, geography, and socioeconomic status (30). CCHS 2004 included individuals of all ages who lived in private dwellings in the 10 Canadian provinces (30). The sample did not include individuals residing in prisons or care facilities, in the Territories, on Indigenous reserves, or full-time members of the Canadian Armed Forces (30). Data collection occurred between January 14, 2004 and January 21, 2005, and respondents were queried about socio-demographic characteristics and health and health-related characteristics (30).

Dietary intake was assessed through interviewer-administered 24-hour recalls conducted using the USDA Automated Multiple-Pass method (AMPM), adapted to the Canadian context (30). The AMPM employs multiple standardized passes to systematically enhance recall of all foods and beverages consumed, including detailed probing for forgotten foods and portion size estimation using visual aids and measurement guides (30). The method has been shown to provide reasonable estimates in short-term studies of group total energy and nutrient intake, including sodium intake at the population level (31–33).

The CCR is a dynamic registry of all new primary cancer case diagnoses since 1992, and is composed of data collected by and reported to Statistics Canada by provincial and territorial cancer registries (34). The registry includes all individuals whose typical place of residence is

Canada, irrespective of whether the individual is part of the provincial/territorial health insurance plan. Cancer classification from 1992 to 2021 is based on the Canadian Council of Cancer Registries standards and reporting requirements (34). From 2021 (year of diagnosis) to present, the International Classification of Diseases for Oncology 3rd edition, 2nd revision (ICD-O-3.2) has been used to classify cancers (34).

4.3.2 HEFI-2019

We chose to model HEFI-2019 total scores rather than component scores, or to split the HEFI-2019 into two separate moderation and adequacy scores. Although there are some advantages to modelling component scores, such as greater insights into component-cancer associations and reducing the inherent challenge of similar aggregate scores despite vastly different dietary patterns, this approach has limitations (35). Analysis of individual components alone relinquishes the ability of indices to account for potentially synergistic effects between dietary components (35). Separating diet quality indices into moderation and adequacy scores attempts to separate healthful and unhealthful components of dietary patterns, which has been found to improve the discriminatory capacity of indices when applying them to understand diet-disease relationships (36). However, given our aim of examining the weighting of component scores and the effect of this weighting approach on the discriminatory capacity of total scores, we chose to model total scores rather than separate moderation and adequacy components to avoid the additional analytical complexity and interpretive challenges that would arise from examining multiple score types simultaneously.

The HEFI-2019 captures alignment with the Canada's Food Guide 2019 healthy food choices recommendations through 10 components (Table 4.1): five food-based components (vegetables and fruits, whole-grain foods, grain foods ratio, protein foods, plant-based protein foods), four nutrient-based components (fatty acids ratio, saturated fats, free sugars, and sodium) and one beverage component (22). All components are scored as ratios, with food-based components quantified using Reference Amounts (RA) – the typical amount of food consumed in a single sitting (in grams) – as numerators and denominators, while nutrient-based components use the amount of respective nutrient or total energy as denominators (22). For food components, ratios were calculated by dividing the total RAs of foods considered within a

specific component by the sum total of RAs for all foods consumed, as well as beverages considered in protein foods (e.g., unsweetened milk, unsweetened plant-based beverages containing protein) (22). For example, the vegetables and fruit component ratio was computed by summing the total RAs of vegetables and fruits consumed and dividing by the total RAs of food consumed, with a perfect score of 20 assigned if the ratio of vegetables and fruits to total foods was ≥ 0.5 (22). Ratios for nutrient-based components were calculated by summing the total amount of nutrient of interest consumed and dividing by the respective denominator (grams of saturated fats for the fatty acids ratio component, and total energy in kcal for the saturated fats, free sugars, and sodium components) (22).

Components are classified as either “adequacy” (higher intakes corresponding to more points) or “moderation” (lower intakes corresponding to more points). Scoring standards are based on the achieved ratio for each component and are based on Canada’s Food Guide recommendations. All components are weighted equally for a maximum of 10 points, except for the vegetables and fruits component, for which a maximum of 20 points is achievable (22). Some dietary constituents are captured across two related components, like grain and whole-grain foods intake, which are captured by the “Whole-grain foods” and “Grain foods ratio” components.

Table 4.1. The Healthy Eating Foods Index 2019 (HEFI-2019) components, description, weighting, and scoring standards adapted from Brassard et. al. 2022 (22)

HEFI-2019	Description	Weighting	Minimum	Maximum
Components		(Maximum component score)	score standard	score standard
Vegetables and fruits	Ratio of total vegetables & fruits to total foods ^a	20	No vegetables or fruits	≥ 0.50

Whole-grain foods	Ratio of whole-grain foods to total foods ^a	5	No whole-grain foods	≥0.25
Grain foods ratio	Ratio of total whole-grain foods to total grain foods ^b	5	No whole-grain foods	1.0 (all grains are whole grains)
Protein foods	Ratio of total protein foods to total foods ^a	5	No protein foods	≥0.25
Plant-based protein foods	Ratio of plant-based protein foods to total protein foods ^c	5	No plant-based protein foods	>0.5
Beverages	Ratio of plain water and unsweetened beverages to total beverages ^d	5	No water and no unsweetened beverages	1.0
Fatty acids ratio	Ratio of Mono- & polyunsaturated fat to total saturated fat	5	≤1.1	≥2.6
Saturated fats	Total saturated fat/energy	5	≥15% of total energy	<10% of total energy
Free sugars	Total free sugars/energy	10	≥20% of total energy	<10% of total energy

Sodium	Total sodium/energy	10	≥ 2.0	< 0.9
--------	---------------------	----	------------	---------

HEFI-2019 Total score = 80

^aTotal foods include all consumed vegetables and fruits, all whole- and non-whole grains foods, protein foods (including processed meats), unsweetened milk and plant-based beverages containing protein, and other foods which are not recommended by Canada's Food Guide 2019 like processed or prepared foods (e.g., sugary breakfast cereals, biscuits and cakes, confectioneries).

^bTotal grain foods include both whole- and non-whole-grain foods.

^cTotal protein foods include fish, shellfish, eggs, poultry, red meat, dairy foods, and plant-based protein foods like legumes, nuts, seeds, and tofu.

^dTotal beverages includes plain water, unsweetened tea/coffee, unsweetened milk and plant-based beverages, and sugary drinks (e.g., soda, sugar-sweetened milks and plant-based beverages, alcohol, and artificially sweetened beverages).

Foods and beverages reported in the 24-hour recalls were coded by Health Canada using the 2005 Canadian Nutrient File (CNF), a recipe file, and a survey foods file (food items not present in the CNF but for which nutritional information was available) (30). The CNF provides the necessary information for the fatty acids ratio (mono and polyunsaturated fat divided by total saturated fats), saturated fats (total saturated fat intake divided by total energy intake), free sugars (denominator of total energy intake only) and sodium (total sodium intake divided by total energy intake) HEFI-2019 components. For the remaining components, it was necessary to access databases from Health Canada that provide information on HEFI-2019 category, RAs, and free sugars values for food codes in the 2015 Canadian Nutrient File (37). Information necessary to compute HEFI-2019 is not provided for all foods included in the 2015 CNF, but instead, all foods reported by respondents in CCHS 2015. Another file from Health Canada that maps food codes assigned to foods and beverages reported in CCHS 2015 to those reported in CCHS 2004 was used to assign RAs and free sugar values within the 2004 dataset. N.S. conducted manual matching for 399 food codes (6.5% of food codes; 6103 total food codes reported in CCHS 2004) used in CCHS 2004 for which no suitable match in the 2015 CNF existed.

An additional 370 food and recipe codes that had matches to the 2015 CNF did not have corresponding HEFI-2019 categories, RAs, or free sugar values. N.S. completed linkages by manually identifying suitable replacement codes with existing information on HEFI-2019 category, RA, and free sugar values. Replacement codes were identified by comparing food code descriptions for foods reported in 2004 with foods reported in 2015 (with available HEFI-2019 information). For instance, “Caribou, cooked” in 2004 did not have a direct match made by Health Canada to the 2015 HEFI-2019 information file, and was linked with the food code with description “Game meat, Indigenous, caribou (reindeer), meat, cooked”. For some food codes, recipe codes were the most suitable linkage, and recipe codes were retrieved from the Government of Canada Open Government Healthy Eating Food Index 2019 web portal (37). S.I.K. reviewed these linkages to ensure suitable replacements were made.

4.3.3 Event and time-to-event

Primary cancer incidence was derived from the CCR according to ICD-9 codes 140-149, 150-159, 160-161 (38). These codes reflect both gastrointestinal and aerodigestive tract cancers. While dietary risk factors may vary somewhat across these anatomical sites, they were analyzed as a combined outcome for two reasons: (i) many dietary factors, such as red and processed meats and vegetables and fruits, as well as other risk factors such as adiposity, alcohol use, and smoking, show consistent associations across the multiple sites, and (ii) statistical power considerations, as site-specific analyses would substantially reduce case numbers for individual cancer types (8). The CCR variable “Cancer case of person, type” and classification “Primary” was used to ascertain a new case of CRC during the reference period (38). Time-to-event was derived from the date of administration of the 24-hr recall and the date of diagnosis, in days, obtained from the CCHS 2004 and CCR, respectively. Participants who did not develop cancer during the follow-up period were censored on December 31, 2016, the last date of cancer registry linkage available.

4.3.4 Covariates

Demographic covariates included sex, age (continuous, in years), educational attainment, total annual household income, and marital status, as well as lifestyle covariates such as smoking status, continuous body mass index (BMI), and alcohol consumption were

included based on prior research indicating the potential correlation between these individual-level factors and the risk of gastrointestinal and aerodigestive cancers and diet quality (39–49). Figure 4.1 presents a directed acyclic graph that proposes the hypothesized causal structure of the exposure, diet quality, and outcome, cancer incidence, as well as the identified covariates.

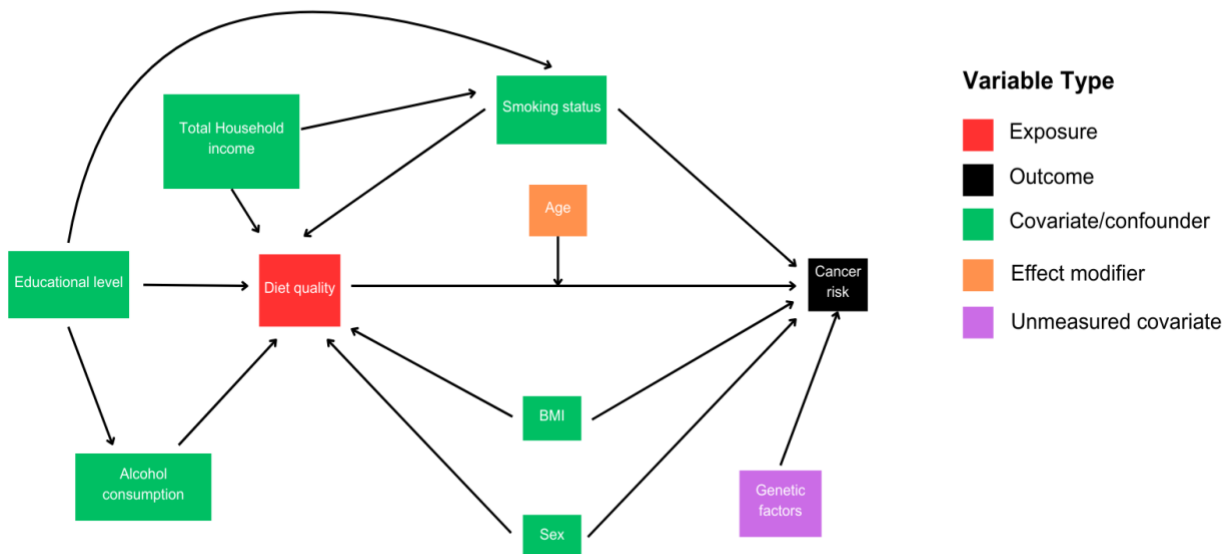


Figure 4.1. Directed acyclic graph illustrating the hypothesized relationships between diet quality and gastrointestinal and aerodigestive cancers, including covariates, effect modifiers, and unmeasured factors.

Educational attainment was derived by Statistics Canada from CCHS 2004 data and included four categories: ‘Less than secondary (high) school graduation (highest)’, ‘Secondary (high) school diploma or equivalent (highest attainment)’, ‘Some post-secondary education (highest)’, and ‘Post-secondary certificate, diploma or degree (highest)’ (50). Income was ascertained using a Statistics Canada-derived variable from CCHS 2004 that categorized total household income into five groups adjusted for household size: ‘Lowest income group’, ‘Lower middle-income group’, ‘Middle income group’, ‘Upper middle-income group’, and ‘Highest income group’, with income thresholds ranging from < \$10,000 to ≥ \$80,000 depending on the number of people in the household. Marital status used a Statistics Canada-derived variable from CCHS 2004 that grouped original responses into four categories (married, common-law,

widowed/divorced/separated, single), which were further recoded into 'Married' and 'Not married' for analysis. Smoking status was retrieved from CCHS 2004 and coded by Statistics Canada as "current smoker", "former smoker", or "never smoker". Alcohol consumption was categorized based on the CCHS 2004 frequency of drinking variable according to Canada's low-risk drinking guidelines. Participants who reported no drinks in the past 12 months were classified as 'No risk drinking'. Those reporting drinking less than once a month to 2-3 times per week were classified as 'Low risk'. Participants drinking 4-6 times per week were classified as 'Moderate risk', and those drinking every day were classified as 'High risk'.

4.3.5 Statistical analyses

Analyses were conducted using RStudio, version 4.0.4, within the SouthWestern Ontario Research Data Centre (SWORDC) at the University of Waterloo. Descriptive statistics for participant characteristics were generated using the "survey" package in R, which accounts for the complex sampling design and provides bootstrap variance estimates (51). For categorical variables, weighted frequencies and percentages were calculated, while continuous variables were summarized using weighted means and standard deviations.

A Cox model with the dependent variable of cancer incidence and the independent variable of *a priori* HEFI-2019 total score, with covariates including sex, age at baseline, educational level, total household income, marital status, smoking status, BMI, and alcohol consumption, was run. The discriminatory power of the model was assessed using Harrell's concordance index (Harrell's C-index). Harrell's C-index describes how well the predicted risk describes the observed sequence of events (52). The "coxph" function within the "survival" package in R was used to run all Cox models (53).

A Cox proportional-hazards model with incident gastrointestinal and aerodigestive cancer as the outcome and the ten HEFI-2019 component scores as primary exposures was then fit. Sex, age, educational attainment, total household income, smoking status, alcohol use and BMI were included as covariates. To identify the components most relevant to cancer risk in this sample, we applied a ridge-penalized likelihood. Prior to applying the ridge penalty, all HEFI-2019 component scores and continuous covariates were standardized to have a mean of zero and a standard deviation of one to ensure comparability of coefficients and prevent scale-

related bias in penalization. To limit over-fitting and allow direct comparison with the unpenalized model, we performed ten random 80/20 training-test splits. For each of the ten training sets, the optimal penalization parameter (λ) was determined independently using 10-fold cross-validation through the “cv.glmnet” function in the glmnet R package, which selects the λ value that minimizes cross-validation error (54,55). For each iteration, the model was estimated in the training set, the penalized β -coefficients for the component scores were extracted, and their absolute values were rescaled to create cancer-specific weights that replaced the standard HEFI-2019 weights. The reweighted HEFI-2019 total score was then used as the dependent variable in a new Cox model in the corresponding test set, from which hazard ratios (HRs) and Harrell’s C-index were obtained. HRs and C-indices from the ten repetitions were pooled and compared with those from the unmodified HEFI-2019 model. Differences in discriminatory capacity were evaluated by examining the magnitude of differences and 95% confidence intervals for the C-indices. Coefficients for covariates in the multivariable models were not interpreted due to the table 2 fallacy (56). These variables were included solely for confounding control rather than as exposures of interest, and their adjusted coefficients do not represent meaningful causal estimates of respective total effects due to mutual adjustment (56).

All estimates are weighted using unique population expansion weights created for the CCHS 2004 – CCR linked dataset, unless otherwise specified. Weighted descriptive results are rounded to the nearest 1,000 according to Statistics Canada vetting requirements for the CCHS 2004 – CCR linked dataset. Sample weights in CCHS 2004 were developed to produce population-representative estimates (30). Initial household weights reflected the inverse probability of selection from either area or telephone sampling frames (30). These weights were adjusted for design features, nonresponse using response-homogeneity groups, and integration across frames (30). Person-level weights were derived by accounting for within-household selection probabilities, followed by winsorization and calibration to known population totals by age, sex, health region, and collection period (30). Bootstrap variance estimation was used to calculate standard errors for all estimates, ensuring that confidence

intervals and statistical inferences appropriately reflected the uncertainty associated with the complex survey design.

Height and weight were not measured for approximately 40% of the CCHS 2004 sample (30). Reasons for missing height and weight information included refusal to be measured, issues with measurement equipment, respondent height exceeding interviewer height, and interviews conducted by telephone (30). BMI could not be imputed as reasons for missing height and weight could not be classified as missing at random or missing completely at random. This is a well-established limitation of CCHS 2004. A unique set of weights is provided by Health Canada to be used in all analyses including measured height and weight or BMI based on measured anthropometrics. All analyses used BMI-specific sample weights.

4.4 Results

Of the 35,017 participants in the complete CCHS 2004 sample, 29,065 participants consented to linkage with other datasets, including the CCR, forming the linked dataset used in this analysis. After applying exclusion criteria, the final analytic sample included 10,530 participants, representing approximately 23,545,000 adults in Canada. The primary exclusions were participants under 18 years of age ($n=11,790$), those with missing height and weight ($n=6,560$), and participants with missing alcohol consumption data due to refusal or not stated responses ($n=50$). Additional exclusions included missing education data ($n=70$), zero energy intake ($n=5$), invalid 24-hour dietary recall data as determined by Health Canada ($n=15$), and cancer events occurring within one year of the dietary recall date ($n=105$).

The analytic sample included 23,545,000 weighted participants, of whom 383,000 developed incident gastrointestinal and aerodigestive cancers during the follow-up period (2005-2016). Participants who developed cancer had a higher mean HEFI-2019 score (44.7, SD 11.2) compared to those who remained cancer-free during follow-up (40.8, SD 11.6). The sample characteristics varied across demographic and lifestyle factors (see Table 4.2). HEFI-2019 component scores stratified by key characteristics (sex, age, educational attainment, income, marital status, smoking status, BMI, and alcohol consumption) are presented in Supplemental Table 1.

Table 4.2. Baseline characteristics of participants by cancer case status, Canadian Community Health Survey–Nutrition 2004 (CCHS 2004)–Canadian Cancer Registry linked cohort (N=23,545,000)+

Variable	Censored observations N (%)	Case N (%)
<i>Sex</i>		
Male	11,458,000 (49.5)	167,000 (43.6)
Female	11,704,000 (50.5)	216,000 (56.4)
<i>Educational attainment</i>		
< High school grad	4,207,000 (18.2)	160,000 (41.8)
High school grad	4,395,000 (19.0)	80,000 (20.9)
Some post-secondary	2,295,000 (9.9)	6000 (1.6)
Post-secondary grad	1,226,500 (53.0)	13,7000 (35.8)
<i>Household Income</i>		
Lowest income group	662,000 (2.9)	9000 (2.3)
Lower middle-income group	126,5000 (5.5)	22,000 (5.7)
Middle-income group	4,360,000 (18.8)	92,000 (24.0)
Upper middle-income group	7,762,000 (33.5)	129,000 (33.7)
Highest income group	7,416,000 (32.0)	100,000 (26.1)
<i>Marital status</i>		
Married	14,914,000 (64.4)	240,000 (62.7)
Not married	8,247,000 (35.6)	143,000 (37.3)
<i>Smoking status</i>		
Current smoker	5,990,000 (25.9)	61,000 (15.9)
Former smoker	6,334,000 (27.3)	154,000 (40.2)
Never smoker	10,837,000 (46.8)	168,000 (43.9)
<i>Alcohol Consumption</i>		
No risk drinking	4,034,000 (17.4)	118,000 (30.8)
Low risk	16,003,000 (69.1)	212,000 (55.4)
Moderate risk	1,343,000 (5.8)	10,000 (2.6)
High Risk	1,781,000 (7.7)	43,000 (11.2)

HEFI 2019 total score ^{††}	40.8 (11.6)	44.7 (11.2)
Baseline age ^{††}	45.2 (17.1)	63.5 (13.8)
Observation time (days) ^{††}	4566.2 (92.4)	2320.7 (1317.3)
BMI ^{††}	27.0 (5.8)	27.9 (5.1)

[†]Weighted estimates using CCHS 2004-CCR linked weights represent the Canadian population. Counts rounded to the nearest 1,000 per Statistics Canada vetting requirements.

^{††}Mean (SD)

The reweighting process greatly altered component weights (Table 4.3), with notable increases for protein foods (from 5 to 16.4), plant-based protein foods (from 5 to 10.7), and free sugars (from 10 to 16.8), and decreases for vegetables and fruits (from 20 to 3.73), whole-grain foods (from 5 to 2.95), and fatty acids ratio (from 5 to 2.06). The sodium component retained its original weight (10.0). Components contributing most to the reweighted score included protein foods, plant-based protein foods, and free sugars.

Table 4.3. Standard and Data-Driven Weights for HEFI-2019 Components Used to Assess Diet Quality in Relation to Gastrointestinal and Aerodigestive Cancer Risk

HEFI-2019 Category	Standard weights	Data-driven weights [†]
Vegetables and Fruits	20	3.73 (3.5)
Whole-grain foods	5	2.95 (2.5)
Grain foods ratio	5	4.57 (1.9)
Protein foods	5	16.4 (3.1)
Plant-based protein foods	5	10.7 (4.1)
Beverages	10	9.3 (6.4)
Fatty acids ratio	5	2.06 (1.5)
Saturated fats	5	3.54 (2.9)
Free sugars	10	16.8 (3.0)
Sodium	10	10.0 (3.8)

[†]Data-driven weights are presented as the mean (SD) across the 10 training sets.

No associations with cancer risk were observed in either model (Table 4.4). Complete pooled model output, including all covariates, is presented in Supplemental Table 2. Model discrimination, as measured by Harrell's C-index, was 0.81 (95% CI: 0.77, 0.85) for the standard

HEFI-2019 score and 0.87 (95% CI: 0.80, 0.93) for the reweighted score, with overlapping confidence intervals indicating no significant difference in discriminatory capacity.

Table 4.4. Association of Standard and Data-Driven Weighted HEFI-2019 Scores with Gastrointestinal and Aerodigestive Cancer Risk in the CCHS 2004–CCR Linked Cohort (N=2,354,400)[†]

	Adj. HR (95% CI)	p-value	Harrell’s C-index (95% CI)
Standard HEFI 2019 total score	1.01 (0.99, 1.04)	0.17	0.81 (0.77, 0.85)
Modified weighting HEFI 2019 total score	1.00 (0.98, 1.02)	0.99	0.87 (0.80, 0.93)

[†]Hazard ratios from weighted Cox proportional hazards models adjusted for age, sex, educational attainment, income, marital status, smoking status, alcohol consumption, and body mass index. Models were weighted using CCHS 2004–CCR linked survey weights with bootstrap variance estimation. Harrell’s C-index represents model discrimination.

4.5 Discussion

In this cohort of adults residing in the Canadian provinces who consented to CCR linkage, we found no evidence of an association between total HEFI-2019 scores and incident gastrointestinal and aerodigestive cancers, using either the *a priori* or the reweighted scores. Reweighting the HEFI-2019 components via a data-driven ridge regression approach yielded only minimal changes in hazard ratios, and, although discrimination (Harrell’s C-index) improved modestly for the modified index, confidence overlapped those of the *a priori* score.

In line with our findings, in an investigation among over 70,000 adults, Moazzen and colleagues (2021) (57) using a single food frequency questionnaire found no association between total scores on the Dutch Dietary Guidelines Index, The Lifelines Diet Score, The American Cancer Society Index, and The World Cancer Research Fund/American Institute for Cancer Research Index, adjusting for lifestyle factors (e.g., age, family history of cancer, education level) and colorectal or upper gastrointestinal cancer risk (57). The Dutch Dietary Guidelines Index includes red meat and processed meat as moderation components—factors directly linked to cancer risk but not explicitly addressed in the HEFI-2019—while the Lifelines Diet Score uses 12 food groups with quintile-based scoring (57,58). Both indices employ equal

weighting across components, which suggests that equal weighting of components may dilute the discriminatory capacity of diet quality indices when examining the association between diet and cancer risk (57,58). Although the American Cancer Society Index and the World Cancer Research Fund/American Institute for Cancer Research Index capture information about diet, adherence to dietary guidance is not the primary focus of these indices, making it difficult to compare their construction to that of the HEFI-2019 (57).

In contrast to the findings of Moazzen and colleagues 2021 (57), Solbak and colleagues (2017) (59) using a single food frequency questionnaire found that fourth quartile as compared to first quartile Healthy Eating Index-2005 Canada scores were associated with decreased risk of cancers of any site (except non-melanoma skin cancer) only in men, a part of the Alberta's Tomorrow Project cohort. This difference in findings from our current investigation may be partially explained by index construction differences between the HEFI-2019 and the HEI-2005 Canada. HEI-2005 Canada includes a broader range of adequacy and moderation components such as dark green and orange vegetables, as well as a 20 point "other foods" component (while all other components were weighted at 5 or 10 points) that explicitly captures foods not recommended in Canada's Food Guide 2019 such as sweets and high-fat snacks, as well as alcohol (60). Although the HEFI-2019 captures similar unhealthy foods by including them in the denominator for food-based component score calculations, explicit inclusion of "other foods" in the HEI-2005 Canada may enhance its ability to capture variance in intakes of foods that may be risk factors for cancer risk. Beyond index content, additional differences likely contributed, including dietary assessment methods (FFQ vs. 24-hour recall), outcome definitions (any cancer vs. site-specific gastrointestinal and aerodigestive cancers), follow-up time and case accrual, and possible effect modification by sex, as the protective association was observed only among men.

Further, a systematic review of epidemiological studies assessed 64 studies investigating diet quality (55 different diet quality indices) and cancer risk and found that 21 studies reveal a positive relationship between diet quality and cancer risk (61). Within this review, 11 studies investigated the association between diet quality and cancers of any site, and only three studies reported that higher diet quality scores were associated with decreased

risk, after adjusting for covariates (61). One of the three studies utilized the Alternate Healthy Eating Index-2010 to characterize dietary patterns, in which all components are weighted equally, but in contrast to the HEFI-2019, the Alternate Healthy Eating Index-2010 included individual components for red and processed meat, as well as alcohol, which may explain the greater discriminatory capacity of the index (62). The remaining two studies to report an association between higher diet quality and decreased risk of any cancer used a 9-component dietary index developed to assess adherence to the Mediterranean dietary pattern (63,64). The 9-component dietary index was weighted equally, and in contrast to the HEFI-2019, explicitly penalized individuals for meat, meat products, and alcohol consumption (63,64). Collectively, these heterogeneous findings across different indices highlight how index construction and weighting frameworks shape observed associations with cancer risk.

We hypothesized that a HEFI-2019 score reweighted using data-driven weights for components may offer stronger associations between diet quality and gastrointestinal and aerodigestive cancer risk and increased discriminatory power within models. The reweighting process significantly altered component weights, with notable increases for protein foods, plant-based protein foods, and free sugars, and decreases for vegetables and fruits, and whole-grain foods. Despite these adjustments, the modified index did not yield any association with cancer risk. The new component scores also appear to conflict with existing cancer-related dietary guidance. For instance, guidance from the World Cancer Research Fund/American Institute for Cancer highlights the importance of consuming whole grains, vegetables, fruit, and beans, and limiting “fast foods” and other processed foods, red and processed meat, sugar-sweetened beverages, and alcohol (8). This apparent disconnect highlights an important challenge of interpreting data-driven reweighting approaches. Reweighting of components reflects the importance that a particular component plays in cancer risk, which implies that components related to protective factors would be down weighted, while components related to known risk factors would be upweighted. This is in line with our findings where the vegetables and fruits, and whole-grain foods components were down weighted, while components such as protein foods, and free sugars were up weighted. However, although reweighting altered the component weights to emphasize risk factors to a greater extent than

protective factors, this reweighting did not strengthen associations between total reweighted HEFI-2019 scores and cancer risk.

In contrast to our findings, Parker et al. (2024) found that data-driven modified-weight HEI-2015 scores were associated with a 23% to 38% reduction in all-cause mortality risk when comparing individuals in the highest modified HEI-2015 score quintile to those in the lowest quintile (66). Using a data-driven approach, components were optimized through Lasso regression models that selected and shrank HEI-2015 component weights based on their contribution to reducing mortality risk (66). This resulted in down-weighting of components with limited predictive value (e.g., added sugars, and sodium) and up-weighting of components with stronger associations (e.g., total fruits) (66). Differences in our findings and existing literature highlight the complexity in examining associations between overall diet quality and cancer or other health outcomes.

Indices such as HEFI-2019 combine moderation components or so-called “unhealthy foods” with adequacy components or so called “healthy foods” in a single score. Within a diet-disease relationship context, this could prove problematic as diseases such as cancer have distinct dietary risk factors, and discriminatory capacity between individual diets may be limited if such risk factors are not adequately differentiated within the scoring frameworks of indices. An investigation of the discriminatory capacity of diet quality indices across differing contexts revealed that the consumption of healthful food groups was positively associated with the concurrent intake of unhealthful food groups (36). When healthy and unhealthy food consumption patterns are correlated, indices that collapse adequacy and moderation into a single score may mask the underlying dietary pattern (36). For instance, a diet high in diverse nutrient-rich foods, but also high in processed meats, sweets, and sugar-sweetened beverages might achieve a similar total score to a diet prioritizing moderation over adequacy, thus masking critical risk factors for disease like cancer. Ultimately, similar scores for diet quality indices can be achieved in many ways, limiting insights into what aspects of an individual dietary pattern are truly healthful or unhealthful in relation to disease outcomes (35).

Component composition of the HEFI-2019 may also explain the null result observed in our study. For example, although red and processed meats are well-established dietary risk

factors for gastrointestinal and aerodigestive cancers, the HEFI-2019, in accordance with Canada's Food Guide 2019, does not directly account for consumption of these foods (67). Instead, individuals with high consumption of red meat, for instance, would in turn have lower scores for the plant-based protein foods component. Here again, middling scores could potentially fail to account for consumption of foods considered to be dietary risk factors for cancer, reducing the utility of the HEFI-2019 in distinguishing between diets that are protective and those that may increase cancer risk. This blending of protective and harmful dietary elements within a single composite score may mask the influence of key exposures, such as red and processed meat intake, ultimately reducing the index's ability to detect meaningful associations with disease outcomes like cancer. As such, the component composition of the HEFI-2019, although aligned with national dietary guidelines, may not be well-suited for use in investigating diet-cancer relationships.

Scoring standards for the HEFI-2019 (Table 4.1) are also important factors to consider when interpreting shifts in weighting of component scores, and the poor discriminatory capacity of the reweighted HEFI-2019 (35). The index relies on a set of threshold-based scoring cutoffs, where individuals receive the maximum score once they meet a predefined intake criterion, regardless of how much higher their intake may be (22). For example, for the vegetables and fruits component, a maximum of 20 points is awarded if an individual's intake reaches or exceeds 50 percent of total food intake (measured in Reference Amounts) (22). As a result, someone who consumes 75 percent of their intake from vegetables and fruits receives the same score as someone consuming just 50 percent, even though these diets differ substantially (22). This lack of scoring gradation above the upper threshold compresses variability in the data and limits the component's ability to differentiate individuals with truly health-promoting diets from those who only minimally meet dietary guidance.

This issue is also relevant for moderation components, such as sodium and free sugars, where scores are capped once intake falls below a specific benchmark (22). There is no additional scoring benefit for further reductions in intake. For instance, individuals who consume very low amounts of sodium receive the same score as those who just meet the cut-point. These ceiling effects reduce the variation in scores for these components, particularly

among individuals with either highly health-promoting or highly detrimental dietary patterns. Without more continuous and sensitive scoring structures, particularly at the upper and lower ends of intake, both standard and reweighted indices may be limited in their ability to reflect actual differences in diet quality.

Additional factors may also explain the discrepancies between our results and existing literature. Our study focused on gastrointestinal and aerodigestive cancers, which may have distinct dietary risk factors compared to the broader outcomes of all-cause mortality and cancer-mortality examined by Parker et al., (2024) (66) and Prentice et al., (2022) (26). The use of ten 80/20 training-test splits in the ridge regression approach reduced statistical power, particularly given the smaller number of cancer cases in test splits (~40 unweighted cases and ~2000 unweighted censored observations per test split) (68). Limited case numbers can hinder the detection of small effect sizes, especially in data-driven approaches requiring validation splits (68). Our analysis also relied on a single 24-hour dietary recall, a method prone to random and systematic error. With respect to random error, a 24-hour recall is not a reflection of usual intake, and episodically consumed foods may not be captured depending on what the respondent ate on a particular day (69). This phenomenon has pertinent effects on subsequently computed dietary indices, as dietary indices capture adherence to some dietary ideal that should be maintained over time, and not necessarily just on a particular day (70). With respect to systematic error, foods recommended by dietary guidance may be overreported, while foods not encouraged by guidance may be underreported. Systematic error could, for instance, lead to overreporting of foods like vegetables and fruits, resulting in overall higher scores for the vegetables and fruit component for the HEFI-2019 across the sample. This would reduce variability in the component's scores across individuals, leading to its down-weighting during the ridge penalization process, which assigns lower importance to variables that contribute less to explaining the exposure–outcome relationship.

Together, our findings provide insights into the performance of diet-quality indices in epidemiological research, though additional studies across different populations and outcomes are needed before drawing broad conclusions about their utility. Indices such as the HEFI-2019, while helpful in tracking broad adherence to national dietary guidance, may not adequately

capture the specific dietary patterns most predictive of disease risk, particularly when applied in epidemiological studies of outcomes like gastrointestinal and aerodigestive cancers. The absence of association, even when components were reweighted using data-driven methods, raises questions about the extent to which such indices possess sufficient discriminatory capacity to detect subtle yet meaningful differences in dietary exposures relevant to cancer risk. Although no studies to date have examined associations between HEFI-2019 and cancer risk, evidence from other outcomes suggests potential utility in larger cohorts or for diseases more aligned with the index's components. For example, in a prospective analysis of over 136,000 UK Biobank participants free of CVD, diabetes, and cancer at baseline, Brassard et al., (2022) (27), found that higher HEFI-2019 scores were associated with a reduced risk of incident CVD over an 11-year follow-up, with 2,843 events observed. Such contrast may reflect the larger sample size and greater number of events in the CVD study, providing enhanced power to detect associations, or differences in how well HEFI-2019 captures risk factors pertinent to CVD versus specific cancers (27). Additional limiting factors of the discriminatory capacity of diet quality indices include ceiling effects and compressed score ranges inherent in threshold-based scoring systems, which restrict the variability needed for robust prediction and risk stratification. Data-driven weighting approaches such as ridge or Lasso regression offer one strategy to improve index performance. However, their utility may be constrained when underlying scoring frameworks do not capture a full gradient of exposure or when between-person variability is low. Moreover, as demonstrated in our analysis, improvements in model discrimination may be modest when applied to complex, multifactorial diseases such as cancer. From a public health standpoint, these results also prompt reflection on the dual use of diet quality indices. While guideline-based scores provide a clear and actionable benchmark to direct public health efforts and messaging, they may require refinement or supplementation with outcome-specific indices in research contexts where predictive accuracy and discriminatory capacity are a central concern.

This study has several notable strengths. It is the first to apply the HEFI-2019 to the CCHS 2004 data, providing a novel assessment of diet quality in relation to gastrointestinal and aerodigestive cancer risk in a Canadian context. Reweighting HEFI-2019 components using a

data-driven ridge penalization approach also offers methodological contributions to understanding the use of diet quality indices in epidemiological research. Data were drawn from a nationally representative cohort of adults living in Canada's provinces, enhancing the generalizability of findings within a Canadian context. Using the CCHS 2004 linked to the Canadian Cancer Registry ensured high reliability for event status determination.

Limitations of the current study include a significant proportion of the sample with missing measured BMI, which led to a reduction in the overall sample size by approximately ~40%. Although the representativeness of the findings is not reduced due to the use of specialized sample weights provided by Statistics Canada, exclusion of ~40% of the sample reduces the overall statistical power of the models, especially during the train-test phases. The study was also limited to participants from the provinces, excluding individuals from the Territories and other communities, which may limit generalizability to all Canadian populations. Key covariates, including educational attainment, marital status, smoking status, and alcohol consumption, were self-reported, making them subject to recall and social desirability bias. Additionally, household income may not accurately represent individual income. Reliance on self-reported dietary intake data introduces the potential for both random and systematic measurement error, which may attenuate the associations observed. Dietary intake was also only assessed at baseline, which may not be reflective of long-term dietary patterns. Manual linkage was needed for 6.5% of all food codes, which introduces the potential for misclassification bias or inconsistency, affecting the precision of dietary intake data and subsequent HEFI-2019 score calculations. Furthermore, because HEFI-2019 categories, Reference Amounts, and free sugar values were assigned based on foods in the 2015 Canadian Nutrient File and subsequently linked to foods in the 2004 Canadian Nutrient File, these values may not accurately reflect the nutritional composition of foods consumed in 2004. Lastly, given the observational nature of this study, the potential for unmeasured confounding remains despite efforts to adjust for known confounders.

Neither the *a priori* HEFI-2019 nor the reweighted HEFI-2019 total scores were associated with the risk of gastrointestinal and aerodigestive cancers. Despite applying a data-driven ridge penalization approach to improve discriminatory capacity, model performance

remained limited. These findings, along with heterogeneous results in other cohorts, suggest that diet quality indices may face challenges in capturing complex diet-disease relationships. Additional research with different samples, outcomes, and methodological approaches is needed to further understand how diet quality indices can best be leveraged in epidemiological applications. Methodological features such as threshold-based scoring, the potential for limited variability in component scores and subsequent total scores, and the limited penalization of scores for known risk factors may further constrain the utility of such indices in epidemiological research. While tools such as the HEFI-2019 serve public health monitoring and surveillance purposes, careful considerations are required for use in investigations of diet and disease relationships.

4.6 References for Chapter 4

1. Canadian Cancer Statistics Advisory Committee, in collaboration with the Canadian Cancer Society, Statistics Canada and the Public Health Agency of Canada. Canadian Cancer Statistics: A 2024 special report on the economic impact of cancer in Canada. Toronto, ON: Canadian Cancer Society; 2024.
2. Brenner DR, Gillis J, Demers AA, Ellison LF, Billette JM, Zhang SX, et al. Projected estimates of cancer in Canada in 2024. CMAJ. 2024 May 13;196(18):E615–23.

3. Bouvard V, Loomis D, Guyton KZ, Grosse Y, Ghissassi FE, Benbrahim-Tallaa L, et al. Carcinogenicity of consumption of red and processed meat. *Lancet Oncol*. 2015 Dec 1;16(16):1599–600.
4. Larsson SC, Wolk A. Red and processed meat consumption and risk of pancreatic cancer: meta-analysis of prospective studies. *Br J Cancer*. 2012 Jan 31;106(3):603–7.
5. Choi Y, Song S, Song Y, Lee JE. Consumption of red and processed meat and esophageal cancer risk: Meta-analysis. *World J Gastroenterol WJG*. 2013 Feb 21;19(7):1020–9.
6. Poorolajal J, Mohammadi Y, Fattahi-Darghlou M, Almasi-Moghadam F. The association between major gastrointestinal cancers and red and processed meat and fish consumption: A systematic review and meta-analysis of the observational studies. *PLOS ONE*. 2024 Jun 26;19(6):e0305994.
7. Wu B, Yang D, Yang S, Zhang G. Dietary Salt Intake and Gastric Cancer Risk: A Systematic Review and Meta-Analysis. *Front Nutr*. 2021 Dec 8;8:801228.
8. Clinton SK, Giovannucci EL, Hursting SD. The World Cancer Research Fund/American Institute for Cancer Research Third Expert Report on Diet, Nutrition, Physical Activity, and Cancer: Impact and Future Directions. *J Nutr*. 2020 Apr;150(4):663–71.
9. Conway DI. Each portion of fruit or vegetable consumed halves the risk of oral cancer. *Evid Based Dent*. 2007 Mar;8(1):19–20.
10. Zalaquett N, Lidoriki I, Lampou M, Saab J, Hadkhale K, Christophi C, et al. Adherence to the Mediterranean Diet and the Risk of Head and Neck Cancer: A Systematic Review and Meta-Analysis of Case–Control Studies. *Nutrients*. 2025 Jan;17(2):287.
11. Diet N. Physical Activity and Stomach Cancer 2016. Revised 2018 [Internet]. World Cancer Research Fund. Am Inst Cancer. 2018
12. Chow WH, Linet MS, McLaughlin JK, Hsing AW, Chien HT, Blot WJ. Risk factors for small intestine cancer. *Cancer Causes Control CCC*. 1993 Mar;4(2):163–9.
13. Papadimitriou N, Markozannes G, Kannellopoulou A, Critselis E, Alhardan S, Karafousia V, et al. An umbrella review of the evidence associating diet and cancer risk at 11 anatomical sites. *Nat Commun*. 2021 Jul 28;12(1):4579.
14. Zhu Q, Shu L, Zhou F, Chen LP, Feng YL. Adherence to the Mediterranean diet and risk of gastric cancer: a systematic review and dose–response meta-analysis. *Front Nutr*. 2023 Sep 8;10:1259453.
15. Oh H, Kim H, Lee DH, Lee A, Giovannucci EL, Kang SS, et al. Different dietary fibre sources and risks of colorectal cancer and adenoma: a dose–response meta-analysis of prospective studies. *Br J Nutr*. 2019 Sep;122(6):605–15.

16. Alexandria V. A Series of Systematic Reviews on the Relationship Between Dietary Patterns and Health Outcomes. Alex VA Available Online <https://nrs.usda.gov/sites/default/files/2019-06/DietaryPatternsReport-FullFinal2.Pdf>. 2014;
17. Burggraf C, Teuber R, Brosig S, Meier T. Review of a priori dietary quality indices in relation to their construction criteria. *Nutr Rev*. 2018 Oct 1;76(10):747–64.
18. Newby PK, Tucker KL. Empirically Derived Eating Patterns Using Factor or Cluster Analysis: A Review. *Nutr Rev*. 2004 May 1;62(5):177–203.
19. Ocké MC. Evaluation of methodologies for assessing the overall diet: dietary quality scores and dietary pattern analysis. *Proc Nutr Soc*. 2013 May;72(2):191–9.
20. Moeller SM, Reedy J, Millen AE, Dixon LB, Newby PK, Tucker KL, et al. Dietary Patterns: Challenges and Opportunities in Dietary Patterns Research: An Experimental Biology Workshop, April 1, 2006. *J Am Diet Assoc*. 2007 Jul 1;107(7):1233–9.
21. Al Jawaldeh A, Osman D, Tawfik A, World Health Organization. Regional Office for the Eastern Mediterranean. Food and nutrition surveillance systems: technical guide for the development of a food and nutrition surveillance system for countries in the Eastern Mediterranean Region. 2013; Available from: <https://iris.who.int/handle/10665/119995>
22. Brassard D, Elvidge Munene LA, St-Pierre S, Guenther PM, Kirkpatrick SI, Slater J, et al. Development of the Healthy Eating Food Index (HEFI)-2019 measuring adherence to Canada's Food Guide 2019 recommendations on healthy food choices. *Appl Physiol Nutr Metab*. 2022 May;47(5):595–610.
23. Schulz CA, Oluwagbemigun K, Nöthlings U. Advances in dietary pattern analysis in nutritional epidemiology. *Eur J Nutr*. 2021 Dec 1;60(8):4115–30.
24. Petersen JM, Naimi AI, Kirkpatrick SI, Bodnar LM. Equal Weighting of the Healthy Eating Index–2010 Components May Not be Appropriate for Pregnancy. *J Nutr*. 2022 May 31;152(8):1886–94.
25. Viallon V, Freisling H, Matta K, Nannsen AØ, Dahm CC, Tjønneland A, et al. On the use of the healthy lifestyle index to investigate specific disease outcomes. *Sci Rep*. 2024 Jul 15;14(1):16330.
26. Prentice RL, Aragaki AK, Van Horn L, Thomson CA, Tinker LF, Manson JE, et al. Mortality Associated with Healthy Eating Index Components and an Empirical-Scores Healthy Eating Index in a Cohort of Postmenopausal Women. *J Nutr*. 2022 Nov 1;152(11):2493–504.
27. Brassard D, Manikpurage HD, Thériault S, Arsenault BJ, Lamarche B. Greater adherence to the 2019 Canada's Food Guide recommendations on healthy food choices reduces the risk of cardiovascular disease in adults: a prospective analysis of UK Biobank data. *Am J Clin Nutr*. 2022 Sep 16;116(6):1748–58.

28. Exploring the association between the Healthy Eating Food Index-2019 (HEFI-2019), the Canadian Healthy Eating Index 2007 (C-HEI 2007), and health among First Nations adults across Canada [Internet]. [cited 2025 Aug 22].
29. Government of Canada SC. Canadian Population Health Survey data (CCHS Annual and Focus Content) integrated with mortality, hospitalization, historical postal codes, cancer, tax data and Census [Internet]. 2020 [cited 2024 Jul 22]. Available from: <https://www.statcan.gc.ca/en/microdata/data-centres/data/cchs>
30. Canada H. Canadian Community Health Survey, Cycle 2.2, Nutrition (2004): A Guide to Accessing and Interpreting the Data [Health Canada, 2006] [Internet]. 2006 [cited 2024 Sep 12]. Available from: <https://www.canada.ca/en/health-canada/services/food-nutrition/food-nutrition-surveillance/health-nutrition-surveys/canadian-community-health-survey-cchs/canadian-community-health-survey-cycle-2-2-nutrition-2004-guide-accessing-interpreting-data-health-canada-2006.html>
31. Blanton CA, Moshfegh AJ, Baer DJ, Kretsch MJ. The USDA Automated Multiple-Pass Method accurately estimates group total energy and nutrient intake. *J Nutr*. 2006 Oct;136(10):2594–9.
32. Rhodes DG, Murayi T, Clemens JC, Baer DJ, Sebastian RS, Moshfegh AJ. The USDA Automated Multiple-Pass Method accurately assesses population sodium intakes. *Am J Clin Nutr*. 2013 May;97(5):958–64.
33. Moshfegh AJ, Rhodes DG, Baer DJ, Murayi T, Clemens JC, Rumpler WV, et al. The US Department of Agriculture Automated Multiple-Pass Method reduces bias in the collection of energy intakes. *Am J Clin Nutr*. 2008 Aug;88(2):324–32.
34. Canadian Cancer Registry System Guide – 2010 Edition [Internet]. [cited 2024 Jul 22]. Available from: https://www23.statcan.gc.ca/imdb-bmdi/pub/document/3207_D10_T1_V2-eng.htm#a18
35. Kirkpatrick SI, Reedy J, Krebs-Smith SM, Pannucci TE, Subar AF, Wilson MM, et al. Applications of the Healthy Eating Index for Surveillance, Epidemiology, and Intervention Research: Considerations and Caveats. *J Acad Nutr Diet*. 2018 Sep;118(9):1603–21.
36. Hanley-Cook GT, Gie SM, Parraguez JP, Hoogerwerf S, Padula de Quadros V, Balcerzak A, et al. Cross-context equivalence and agreement of healthy diet metrics for national and global monitoring: a multicountry analysis of cross-sectional quantitative 24-hour dietary intake studies. *Am J Clin Nutr*. 2024 Nov 1;120(5):1093–104.
37. Canada H. The Healthy Eating Food Index 2019 - Open Government Portal [Internet]. [cited 2024 Sep 12]. Available from: <https://open.canada.ca/data/en/dataset/29892c85-2ff5-484c-873c-f494ffba6e1b>

38. Government of Canada SC. Canadian Cancer Registry (CCR) [Internet]. 2019 [cited 2024 Sep 12]. Available from:
<https://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&Id=1211637>
39. Mouw T, Koster A, Wright ME, Blank MM, Moore SC, Hollenbeck A, et al. Education and Risk of Cancer in a Large Cohort of Men and Women in the United States. *PLOS ONE*. 2008 Nov 4;3(11):e3639.
40. Koyanagi YN, Ito H, Matsuo K, Sugawara Y, Hidaka A, Sawada N, et al. Smoking and Pancreatic Cancer Incidence: A Pooled Analysis of 10 Population-Based Cohort Studies in Japan. *Cancer Epidemiol Biomarkers Prev*. 2019 Aug 1;28(8):1370–8.
41. Botteri E, Iodice S, Bagnardi V, Raimondi S, Lowenfels AB, Maisonneuve P. Smoking and Colorectal Cancer: A Meta-analysis. *JAMA*. 2008 Dec 17;300(23):2765–78.
42. Scherübl H. Tobacco Smoking and Gastrointestinal Cancer Risk. *Visc Med*. 2022 Mar 17;38(3):217–22.
43. Altieri A, Bosetti C, Talamini R, Gallus S, Franceschi S, Levi F, et al. Cessation of smoking and drinking and the risk of laryngeal cancer. *Br J Cancer*. 2002 Nov;87(11):1227–9.
44. Khalifeh M, Ginex P, Boffetta P. Reduction of head and neck cancer risk following smoking cessation: a systematic review and meta-analysis. *BMJ Open*. 2024 Aug 9;14(8):e074723.
45. Bagnardi V, Rota M, Botteri E, Tramacere I, Islami F, Fedirko V, et al. Alcohol consumption and site-specific cancer risk: a comprehensive dose-response meta-analysis. *Br J Cancer*. 2015 Feb 3;112(3):580–93.
46. Larsson SC, Wolk A. Obesity and colon and rectal cancer risk: a meta-analysis of prospective studies2. *Am J Clin Nutr*. 2007 Sep 1;86(3):556–65.
47. Turati F, Tramacere I, La Vecchia C, Negri E. A meta-analysis of body mass index and esophageal and gastric cardia adenocarcinoma. *Ann Oncol Off J Eur Soc Med Oncol*. 2013 Mar;24(3):609–17.
48. Maisonneuve P, Lowenfels AB. Risk factors for pancreatic cancer: a summary review of meta-analytical studies. *Int J Epidemiol*. 2015 Feb 1;44(1):186–98.
49. Shi X, Deng G, Wen H, Lin A, Wang H, Zhu L, et al. Role of body mass index and weight change in the risk of cancer: A systematic review and meta-analysis of 66 cohort studies. *J Glob Health*. 14:04067.
50. Government of Canada SC. Canadian Community Health Survey - Nutrition (CCHS) [Internet]. 2014 [cited 2024 Sep 11]. Available from:
<https://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&SDDS=5049>

51. Lumley T. Analysis of Complex Survey Samples. *J Stat Softw.* 2004 Apr 15;9:1–19.
52. Newson RB. Comparing the Predictive Powers of Survival Models Using Harrell's C or Somers' D. *Stata J.* 2010 Sep 1;10(3):339–58.
53. CRAN: Package survival [Internet]. [cited 2025 Aug 22]. Available from: <https://cran.r-project.org/web/packages/survival/index.html>
54. Friedman JH, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *J Stat Softw.* 2010 Feb 2;33:1–22.
55. Simon N, Friedman JH, Hastie T, Tibshirani R. Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent. *J Stat Softw.* 2011 Mar 9;39:1–13.
56. Westreich D, Greenland S. The table 2 fallacy: presenting and interpreting confounder and modifier coefficients. *Am J Epidemiol.* 2013 Feb 15;177(4):292–8.
57. Moazzen S, Cortes-Ibañez FO, van der Vegt B, Alizadeh BZ, de Bock GH. Diet quality indices and gastrointestinal cancer risk: results from the Lifelines study. *Eur J Nutr.* 2022;61(1):317–27.
58. Assessment of the Dutch Healthy Diet index 2015 in the Lifelines cohort study at baseline | *European Journal of Clinical Nutrition* [Internet]. [cited 2025 Aug 22]. Available from: <https://www.nature.com/articles/s41430-023-01372-x>
59. Solbak NM, Xu JY, Vena JE, Csizmadi I, Whelan HK, Robson PJ. Diet quality is associated with reduced incidence of cancer and self-reported chronic disease: Observations from Alberta's Tomorrow Project. *Prev Med.* 2017 Aug 1;101:178–87.
60. Garriguet D. Diet quality in Canada. *Health Rep.* 2009 Sep;20(3):41–52.
61. Potter J, Brown L, Williams RL, Byles J, Collins CE. Diet Quality and Cancer Outcomes in Adults: A Systematic Review of Epidemiological Studies. *Int J Mol Sci.* 2016 Jul;17(7):1052.
62. Chiuve SE, Fung TT, Rimm EB, Hu FB, McCullough ML, Wang M, et al. Alternative Dietary Indices Both Strongly Predict Risk of Chronic Disease, .. *J Nutr.* 2012 Jun 1;142(6):1009–18.
63. Couto E, Boffetta P, Lagiou P, Ferrari P, Buckland G, Overvad K, et al. Mediterranean dietary pattern and cancer risk in the EPIC cohort. *Br J Cancer.* 2011 Apr 26;104(9):1493–9.
64. Benetou V, Trichopoulou A, Orfanos P, Naska A, Lagiou P, Boffetta P, et al. Conformity to traditional Mediterranean diet and cancer incidence: the Greek EPIC cohort. *Br J Cancer.* 2008 Jul 8;99(1):191–5.

65. The Elements of Statistical Learning: Data Mining, Inference, and Prediction | SpringerLink [Internet]. [cited 2025 Aug 7]. Available from: <https://link.springer.com/book/10.1007/978-0-387-21606-5>
66. Parker HW, Oaks BM, Buchanan AL, Vadiveloo MK. Modified-Weight Healthy Eating Index-2015 Scores Are More Strongly Associated With Mortality Risk Than Standard Scores. *J Acad Nutr Diet*. 2024 Mar 1;124(3):331–45.
67. Canada H. Welcome to Canada's food guide [Internet]. 2021 [cited 2024 Sep 11]. Available from: <https://food-guide.canada.ca/en/>
68. Steyerberg EW, FRANK E, HARRELL, Regression Modeling Strategies: With Applications, to Linear Models, Logistic and Ordinal Regression, and Survival Analysis, 2nd ed. Heidelberg: Springer. *Biometrics*. 2016 Sep 1;72(3):1006–7.
69. Kirkpatrick SI, Troiano RP, Barrett B, Cunningham C, Subar AF, Park Y, et al. Measurement Error Affecting Web- and Paper-Based Dietary Assessment Instruments: Insights From the Multi-Cohort Eating and Activity Study for Understanding Reporting Error. *Am J Epidemiol*. 2022 Feb 8;191(6):1125–39.
70. Guenther PM, Reedy J, Krebs-Smith SM. Development of the Healthy Eating Index-2005. *J Am Diet Assoc*. 2008 Nov 1;108(11):1896–901.

Chapter 6: General Discussion

6.1 Summary of key findings

This thesis examined the relationship between diet quality, as assessed by the HEFI-2019, and the risk of gastrointestinal and aerodigestive cancers in Canadian adults. Chapter 4, the central manuscript reporting on this work, examined whether observed associations differed when HEFI-2019 scores were calculated using the *a priori* weighting scheme versus a data-driven reweighted version of the index.

Findings from Chapter 4 did not provide evidence of clear associations between HEFI-2019 scores and cancer risk. Both the standard and reweighted indices produced hazard ratios that were close to null, and model discrimination showed only modest differences between the two approaches. These results suggest that, within this sample, overall diet quality as captured by HEFI-2019 may not have been strongly predictive of cancer incidence, though limitations such as case numbers, dietary assessment methods, and the multifactorial nature of cancer must be considered when interpreting these findings.

The reweighting process substantially altered the contribution of individual components, giving greater emphasis to protein foods, plant-based proteins, and free sugars, while reducing the weight of vegetables, fruits, and whole grains. Despite these shifts, the modified index did not appear to improve the prediction of cancer risk. Moreover, the down-weighting of components emphasized in existing dietary guidance raises questions about how well data-driven approaches align with established evidence on diet and cancer. These discrepancies may reflect factors such as measurement error, limited between-person variability in intakes, or the specific characteristics of this cohort.

Taken together, the results highlight the challenges of using broad diet quality indices to study complex outcomes such as cancer. Composite scores that combine both adequacy and moderation components may mask the influence of specific risk factors, while threshold-based scoring can compress variability and limit discriminatory capacity. This work contributes to ongoing discussions about how diet quality indices are best applied in epidemiological research.

6.2 Strengths

This work has several strengths. To the best of our knowledge, this is the first study to apply the HEFI-2019 to the CCHS 2004. The ability to examine this index in an earlier, nationally representative dietary survey extends its application beyond more recent cohorts and offers insights into its performance in relation to cancer outcomes.

The use of a data-driven approach to reweighting to examine whether alternative operationalizations of the HEFI-2019 might yield different insights offers a key methodological contribution. By applying ridge regression to derive empirical weights for index components, this work adds to discussions about how the construction of diet quality indices may influence observed diet-disease relationships. Although regression results were not markedly different from those based on the *a priori* index, the comparison itself adds value by illustrating both the potential and limitations of one strategy to improve the discriminatory capacity of diet quality indices.

The national scope of the CCHS 2004 is another important strength. As a population-based survey designed to capture dietary intake across all provinces, the dataset allows for findings that are broadly representative of adults in Canada at the time of data collection. The application of survey weights and bootstrap variance estimation further supports representativeness and accurate standard error estimates. In addition, the use of the CCR provided prospective insights into diet-cancer relationships in this study, while also ensuring high reliability for determining event status.

6.3 Study limitations

This thesis has several limitations that should be considered when interpreting the findings. One important limitation is the substantial proportion of participants with missing measured body mass index (BMI). Approximately 40% of the analytic sample was excluded due to this missing information, which reduced the overall sample size and decreased statistical power, particularly during train–test phases of the analysis. While the use of specialized sample weights provided by Statistics Canada helps preserve representativeness, the reduction in sample size remains a constraint.

The study population was also limited to individuals residing in the provinces, as the Canadian Community Health Survey (CCHS) 2004–Nutrition did not include participants from the Territories or other communities. This exclusion may limit the generalizability of findings to all populations in Canada.

Several covariates, including educational attainment, marital status, smoking status, and alcohol consumption, were self-reported. As such, they may be subject to recall bias and social desirability bias. Similarly, household income was used as a proxy for socioeconomic status; however, it may not accurately reflect individual-level income, thereby introducing the possibility of misclassification.

Dietary intake was also self-reported and assessed only once at baseline. Recall data are affected by both random error and systematic error, and diet measured at a single time point may not fully represent long-term intake patterns relevant to cancer development. Additional measurement concerns include the manual linkage required for 6.5% of food codes, which may have introduced misclassification or inconsistency. Moreover, HEFI-2019 categories, Reference Amounts, and free sugar values were assigned using the 2015 Canadian Nutrient File and then applied retrospectively to foods reported in the 2004 survey. This process may have introduced inaccuracies if the nutrient composition of foods differed between 2004 and 2015.

Finally, as with any observational study, unmeasured or residual confounding cannot be ruled out. Although known covariates were adjusted for, other factors that were not available in the dataset may have influenced both diet quality and cancer outcomes.

6.4 Implications for future research

Assessing diet-disease relations is complex. In this thesis, the weighting scheme of the HEFI-2019 was examined, which is one avenue of exploring the discriminatory capacity of such indices. As discussed in Chapter 4, component composition and the scoring standards applied are also critical elements of index construction, since they determine how diverse dietary intakes are translated into aggregate scores. These structural choices can substantially influence the degree to which indices capture variation in diet quality and, ultimately, their ability to distinguish between individuals at higher or lower risk of disease. Future work could therefore usefully extend beyond weighting to examine how modifications to component

selection, thresholds, or scoring ranges affect diet-disease associations. Additionally, it may be valuable to investigate how combining adequacy and moderation components within a single score shapes observed relationships, or whether outcome-specific indices tailored to disease-related risk factors yield clearer insights. Future work should also perform appropriate measurement error correction to account for random and systematic error inherent to 24-hour food recall. Together, these considerations highlight the importance of ongoing methodological reflection in nutritional epidemiology and the need to evaluate how different aspects of index construction influence both research findings and their implications for dietary guidance.

Dietary patterns are inherently dynamic and may change considerably over time in response to multiple factors, including age, health status, socioeconomic circumstances, food availability, and broader societal or policy changes. Relying on a single 24-hour recall, therefore, provides only a snapshot of intake, which may not accurately reflect usual or long-term dietary patterns most relevant for chronic disease development. This limitation is particularly important in the context of cancer research, where the latency period between exposure and outcome can span many years or even decades. Future research would benefit from repeated dietary measures within cohorts, which would enable the capture of within-person variability and temporal changes in diet.

This thesis also makes key contributions to dietary assessment capacity in Canada. A key output of this work is the linkage of food codes reported in the CCHS 2004 to food codes in CCHS 2015, with available information to calculate the HEFI-2019. This resource will be made publicly available, enabling other researchers to compute HEFI-2019 scores in the CCHS 2004 cohort. Given that this cohort is linked to multiple health databases (i.e., Discharge Abstract Database, National Ambulatory Care Reporting System, Ontario Mental Health Reporting System, Canadian Vital Statistics Death Database, Historical Postal Codes, Census of Population, and the T1 Family File), the availability of HEFI-2019 scores opens opportunities for a wide range of prospective investigations into diet-disease relationships, particularly those with long latency periods that cannot be examined using more recent CCHS 2015 data.

6.5 Conclusion

This research was the first to apply the HEFI-2019 to the CCHS 2004 linked with the Canadian Cancer Registry. Examining diet quality in relation to incident gastrointestinal and aerodigestive cancers provides new insights into how this index performs in a Canadian cohort with long-term follow-up. The analyses in Chapter 4 suggest that total HEFI-2019 scores, whether based on *a priori* weights or empirically derived weights, were not associated with cancer risk. While the reweighting approach substantially shifted the emphasis of the index, there was little indication that this improved discrimination or altered the overall conclusions. These findings underscore the challenges of capturing complex diet-disease relationships with broad indices and highlight the importance of continued methodological reflection in how dietary exposures are operationalized in nutritional epidemiology.

References

1. Canadian Cancer Statistics Advisory Committee in collaboration with the Canadian Cancer Society, Statistics Canada and the Public Health Agency of Canada. Canadian Cancer Statistics: A 2024 special report on the economic impact of cancer in Canada. Toronto, ON: Canadian Cancer Society; 2024.
2. Bouvard V, Loomis D, Guyton KZ, Grosse Y, Ghissassi FE, Benbrahim-Tallaa L, et al. Carcinogenicity of consumption of red and processed meat. *Lancet Oncol*. 2015 Dec 1;16(16):1599–600.
3. Larsson SC, Wolk A. Red and processed meat consumption and risk of pancreatic cancer: meta-analysis of prospective studies. *Br J Cancer*. 2012 Jan 31;106(3):603–7.
4. Choi Y, Song S, Song Y, Lee JE. Consumption of red and processed meat and esophageal cancer risk: Meta-analysis. *World J Gastroenterol WJG*. 2013 Feb 21;19(7):1020–9.
5. Poorolajal J, Mohammadi Y, Fattahi-Darghlou M, Almasi-Moghadam F. The association between major gastrointestinal cancers and red and processed meat and fish consumption: A systematic review and meta-analysis of the observational studies. *PLOS ONE*. 2024 Jun 26;19(6):e0305994.
6. Clinton SK, Giovannucci EL, Hursting SD. The World Cancer Research Fund/American Institute for Cancer Research Third Expert Report on Diet, Nutrition, Physical Activity, and Cancer: Impact and Future Directions. *J Nutr*. 2020 Apr;150(4):663–71.
7. Conway DI. Each portion of fruit or vegetable consumed halves the risk of oral cancer. *Evid Based Dent*. 2007 Mar;8(1):19–20.
8. Zalaquett N, Lidoriki I, Lampou M, Saab J, Hadkhale K, Christophi C, et al. Adherence to the Mediterranean Diet and the Risk of Head and Neck Cancer: A Systematic Review and Meta-Analysis of Case–Control Studies. *Nutrients*. 2025 Jan;17(2):287.
9. World Cancer Research Fund/American Institute for Cancer Research. Continuous Update Project Expert Report 2018. Diet, nutrition, physical activity and stomach cancer. 2018;
10. Chow WH, Linet MS, McLaughlin JK, Hsing AW, Chien HT, Blot WJ. Risk factors for small intestine cancer. *Cancer Causes Control CCC*. 1993 Mar;4(2):163–9.

11. Papadimitriou N, Markozannes G, Kannelopoulou A, Critselis E, Alhardan S, Karafousia V, et al. An umbrella review of the evidence associating diet and cancer risk at 11 anatomical sites. *Nat Commun*. 2021 Jul 28;12(1):4579.
12. Zhu Q, Shu L, Zhou F, Chen LP, Feng YL. Adherence to the Mediterranean diet and risk of gastric cancer: a systematic review and dose–response meta-analysis. *Front Nutr*. 2023 Sep 8;10:1259453.
13. Alexandria V. A Series of Systematic Reviews on the Relationship Between Dietary Patterns and Health Outcomes. Alex VA Available Online <https://www.usda.gov/sites/default/files/2019-06/DietaryPatternsReport-FullFinal2.Pdf>. 2014;
14. Ocké MC. Evaluation of methodologies for assessing the overall diet: dietary quality scores and dietary pattern analysis. *Proc Nutr Soc*. 2013 May;72(2):191–9.
15. Moeller SM, Reedy J, Millen AE, Dixon LB, Newby PK, Tucker KL, et al. Dietary Patterns: Challenges and Opportunities in Dietary Patterns Research: An Experimental Biology Workshop, April 1, 2006. *J Am Diet Assoc*. 2007 Jul 1;107(7):1233–9.
16. Al Jawaldeh A, Osman D, Tawfik A, World Health Organization. Regional Office for the Eastern Mediterranean. Food and nutrition surveillance systems: technical guide for the development of a food and nutrition surveillance system for countries in the Eastern Mediterranean Region. 2013; Available from: <https://iris.who.int/handle/10665/119995>
17. Brassard D, Elvidge Munene LA, St-Pierre S, Guenther PM, Kirkpatrick SI, Slater J, et al. Development of the Healthy Eating Food Index (HEFI)-2019 measuring adherence to Canada’s Food Guide 2019 recommendations on healthy food choices. *Appl Physiol Nutr Metab*. 2022 May;47(5):595–610.
18. Schulz CA, Oluwagbemigun K, Nöthlings U. Advances in dietary pattern analysis in nutritional epidemiology. *Eur J Nutr*. 2021 Dec 1;60(8):4115–30.
19. Petersen JM, Naimi AI, Kirkpatrick SI, Bodnar LM. Equal Weighting of the Healthy Eating Index–2010 Components May Not be Appropriate for Pregnancy. *J Nutr*. 2022 May 31;152(8):1886–94.
20. Viallon V, Freisling H, Matta K, Nannsen AØ, Dahm CC, Tjønneland A, et al. On the use of the healthy lifestyle index to investigate specific disease outcomes. *Sci Rep*. 2024 Jul 15;14(1):16330.
21. Prentice RL, Aragaki AK, Van Horn L, Thomson CA, Tinker LF, Manson JE, et al. Mortality Associated with Healthy Eating Index Components and an Empirical-Scores Healthy Eating Index in a Cohort of Postmenopausal Women. *J Nutr*. 2022 Nov 1;152(11):2493–504.
22. Forouzanfar MH, Alexander L, Anderson HR, Bachman VF, Biryukov S, Brauer M, et al. Global, regional, and national comparative risk assessment of 79 behavioural,

- environmental and occupational, and metabolic risks or clusters of risks in 188 countries, 1990–2013: a systematic analysis for the Global Burden of Disease Study 2013. *The Lancet*. 2015 Dec 5;386(10010):2287–323.
23. GBD 2016 Risk Factors Collaborators. Global, regional, and national comparative risk assessment of 84 behavioural, environmental and occupational, and metabolic risks or clusters of risks, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016. *Lancet Lond Engl*. 2017 Sep 16;390(10100):1345–422.
 24. Gropper SS. The Role of Nutrition in Chronic Disease. *Nutrients*. 2023 Jan 28;15(3):664.
 25. Taylor RM, Haslam RL, Herbert J, Whatnall MC, Trijsburg L, de Vries JHM, et al. Diet quality and cardiovascular outcomes: A systematic review and meta-analysis of cohort studies. *Nutr Diet J Dietit Assoc Aust*. 2024 Feb;81(1):35–50.
 26. Grosso G, Bella F, Godos J, Sciacca S, Del Rio D, Ray S, et al. Possible role of diet in cancer: systematic review and multiple meta-analyses of dietary patterns, lifestyle factors, and cancer risk. *Nutr Rev*. 2017 Jun 1;75(6):405–19.
 27. Liefvers JRL, Ekwaru JP, Ohinmaa A, Veugelers PJ. The economic burden of not meeting food recommendations in Canada: The cost of doing nothing. *PLoS ONE* [Internet]. 2018 [cited 2024 Sep 11];13(4). Available from: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5922758/>
 28. World Health Organization, FAO, editors. Diet, nutrition, and the prevention of chronic diseases: report of a WHO-FAO Expert Consultation ; [Joint WHO-FAO Expert Consultation on Diet, Nutrition, and the Prevention of Chronic Diseases, 2002, Geneva, Switzerland]. Geneva: World Health Organization; 2003. 149 p. (WHO technical report series).
 29. Hu FB. Dietary pattern analysis: a new direction in nutritional epidemiology. *Curr Opin Lipidol*. 2002 Feb;13(1):3.
 30. Cespedes EM, Hu FB. Dietary patterns: from nutritional epidemiologic analysis to national guidelines. *Am J Clin Nutr*. 2015 May;101(5):899–900.
 31. Gonzalez Fischer C, Garnett T. Plates, pyramids, planet. Developments in national healthy and sustainable dietary guidelines: a state of play assessment. 2016.
 32. Canada H. Welcome to Canada’s food guide [Internet]. 2021 [cited 2024 Sep 11]. Available from: <https://food-guide.canada.ca/en/>
 33. NHANES - About the National Health and Nutrition Examination Survey [Internet]. 2024 [cited 2024 Sep 11]. Available from: https://www.cdc.gov/nchs/nhanes/about_nhanes.htm

34. Government of Canada SC. Canadian Community Health Survey - Nutrition (CCHS) [Internet]. 2014 [cited 2024 Sep 11]. Available from: <https://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&SDDS=5049>
35. Herrick KA, Ogden CL. Chapter 12 - Nutrition surveillance. In: Marriott BP, Birt DF, Stallings VA, Yates AA, editors. *Present Knowledge in Nutrition (Eleventh Edition)* [Internet]. Academic Press; 2020 [cited 2024 Sep 11]. p. 217–33. Available from: <https://www.sciencedirect.com/science/article/pii/B9780128184608000125>
36. 24-hour Dietary Recall (24HR) At a Glance | Dietary Assessment Primer [Internet]. [cited 2024 Sep 11]. Available from: <https://dietassessmentprimer.cancer.gov/profiles/recall/>
37. Cohen HW, Hailpern SM, Fang J, Alderman MH. Sodium Intake and Mortality in the NHANES II Follow-up Study. *Am J Med*. 2006 Mar 1;119(3):275.e7-275.e14.
38. Burggraf C, Teuber R, Brosig S, Meier T. Review of a priori dietary quality indices in relation to their construction criteria. *Nutr Rev*. 2018 Oct 1;76(10):747–64.
39. Newby PK, Tucker KL. Empirically Derived Eating Patterns Using Factor or Cluster Analysis: A Review. *Nutr Rev*. 2004 May 1;62(5):177–203.
40. Kim S, Haines PS, Siega-Riz AM, Popkin BM. The Diet Quality Index-International (DQI-I) Provides an Effective Tool for Cross-National Comparison of Diet Quality as Illustrated by China and the United States. *J Nutr*. 2003 Nov 1;133(11):3476–84.
41. Guenther PM, Reedy J, Krebs-Smith SM. Development of the Healthy Eating Index-2005. *J Am Diet Assoc*. 2008 Nov 1;108(11):1896–901.
42. Rumawas ME, Dwyer JT, Mckeown NM, Meigs JB, Rogers G, Jacques PF. The Development of the Mediterranean-Style Dietary Pattern Score and Its Application to the American Diet in the Framingham Offspring Cohort. *J Nutr*. 2009 Jun;139(6):1150–6.
43. Krebs-Smith SM, Pannucci TE, Subar AF, Kirkpatrick SI, Lerman JL, Tooze JA, et al. Update of the Healthy Eating Index: HEI-2015. *J Acad Nutr Diet*. 2018 Sep 1;118(9):1591–602.
44. Kirkpatrick SI, Reedy J, Krebs-Smith SM, Pannucci TE, Subar AF, Wilson MM, et al. Applications of the Healthy Eating Index for Surveillance, Epidemiology, and Intervention Research: Considerations and Caveats. *J Acad Nutr Diet*. 2018 Sep;118(9):1603–21.
45. Parker HW, Oaks BM, Buchanan AL, Vadiveloo MK. Modified-Weight Healthy Eating Index-2015 Scores Are More Strongly Associated With Mortality Risk Than Standard Scores. *J Acad Nutr Diet*. 2024 Mar 1;124(3):331–45.
46. Brenner DR, Gillis J, Demers AA, Ellison LF, Billette JM, Zhang SX, et al. Projected estimates of cancer in Canada in 2024. *CMAJ*. 2024 May 13;196(18):E615–23.

47. Wu B, Yang D, Yang S, Zhang G. Dietary Salt Intake and Gastric Cancer Risk: A Systematic Review and Meta-Analysis. *Front Nutr*. 2021 Dec 8;8:801228.
48. Freedman ND, Park Y, Subar AF, Hollenbeck AR, Leitzmann MF, Schatzkin A, et al. Fruit and vegetable intake and head and neck cancer risk in a large United States prospective cohort study. *Int J Cancer*. 2008 May 15;122(10):2330–6.
49. World Cancer Research Fund/ American Institute for Cancer Research. Continuous Update Project Expert Report 2018. Diet, nutrition, physical activity, and cancers of the mouth, pharynx and larynx.
50. Pavia M, Pileggi C, Nobile CGA, Angelillo IF. Association between fruit and vegetable consumption and oral cancer: a meta-analysis of observational studies. *Am J Clin Nutr*. 2006 May;83(5):1126–34.
51. Aune D, Chan DSM, Lau R, Vieira R, Greenwood DC, Kampman E, et al. Dietary fibre, whole grains, and risk of colorectal cancer: systematic review and dose-response meta-analysis of prospective studies. *The BMJ*. 2011 Nov 10;343:d6617.
52. Vieira AR, Abar L, Chan DSM, Vingeliene S, Polemiti E, Stevens C, et al. Foods and beverages and colorectal cancer risk: a systematic review and meta-analysis of cohort studies, an update of the evidence of the WCRF-AICR Continuous Update Project. *Ann Oncol Off J Eur Soc Med Oncol*. 2017 Aug 1;28(8):1788–802.

Appendix A: Supplementary materials for Chapter 4

Supplemental Table 1. Mean (Standard Deviation) of HEFI-2019 Component Scores by Sociodemographic and Lifestyle Characteristics, Canadian Community Health Survey–Nutrition 2004 (CCHS 2.2)–Canadian Cancer Registry Linked Cohort (N = 2,354,400)

HEFI-2019 Components (mean, SD)	Sex		Educational attainment				Income					Marital status		Smoking status			Alcohol Consumption			
	Male	Female	< High school grad	High school grad	Some post- secondary	Post- secondary grad	Lowest income	Lower middle- income	Middle income	Upper middle- income	Highest income	Married	Not married	Current	Former	Never	No risk	Low risk	Moderate risk	High Risk
Vegetables and Fruits	8.6 (0.2)	10.2 (0.1)	9.6 (0.2)	8.9 (0.3)	8.4 (0.3)	9.7 (0.2)	9.8 (0.7)	9.5 (0.4)	9.7 (0.3)	9.2 (0.2)	9.4 (0.2)	9.7 (0.2)	8.9 (0.2)	7.8 (0.2)	9.9 (0.2)	10 (0.2)	10.1 (0.3)	9.2 (0.1)	9.1 (0.4)	9.8 (0.4)
Whole-grain foods	0.9 (0)	1.1 (0)	0.9 (0)	0.8 (0.1)	1 (0.1)	1.1 (0)	1.2 (0.3)	1 (0.2)	0.9 (0.1)	1 (0)	1 (0)	1 (0)	0.9 (0)	0.7 (0)	1 (0.1)	1.1 (0)	1.1 (0.1)	1 (0)	1 (0.1)	0.9 (0.1)
Grain foods ratio	1.1 (0)	1.4 (0)	1.1 (0.1)	1.1 (0.1)	1.3 (0.1)	1.3 (0)	1.3 (0.2)	1.2 (0.1)	1.2 (0.1)	1.3 (0.1)	1.3 (0.1)	1.3 (0)	1.1 (0)	0.9 (0.1)	1.3 (0.1)	1.4 (0)	1.3 (0.1)	1.2 (0)	1.2 (0.1)	1.2 (0.1)
Protein foods	3.9 (0)	3.8 (0)	3.6 (0.1)	3.7 (0.1)	3.9 (0.1)	3.9 (0)	3.6 (0.1)	3.7 (0.1)	3.7 (0.1)	3.9 (0)	3.9 (0)	3.8 (0)	3.8 (0)	3.7 (0.1)	3.8 (0)	3.9 (0)	3.6 (0.1)	3.8 (0)	3.9 (0.1)	4 (0.1)
Plant-based protein foods	0.8 (0)	0.9 (0)	0.8 (0.1)	0.7 (0.1)	0.7 (0.1)	1 (0)	0.6 (0.1)	1 (0.2)	0.8 (0.1)	0.9 (0.1)	0.9 (0.1)	0.9 (0)	0.7 (0)	0.7 (0.1)	0.9 (0.1)	1 (0)	0.9 (0.1)	0.9 (0)	0.8 (0.1)	0.9 (0.1)
Beverages	6.7 (0.1)	7.7 (0.1)	7.3 (0.1)	7.3 (0.1)	6.9 (0.2)	7.2 (0.1)	6.9 (0.3)	7.5 (0.2)	7.5 (0.1)	7.2 (0.1)	7.1 (0.1)	7.4 (0.1)	6.9 (0.1)	6.8 (0.1)	7.5 (0.1)	7.3 (0.1)	7.9 (0.1)	7.2 (0.1)	6.5 (0.2)	6.2 (0.2)
Fatty acids ratio	2.5 (0.1)	2.5 (0)	2.7 (0.1)	2.4 (0.1)	2.5 (0.1)	2.5 (0)	2.5 (0.2)	2.9 (0.1)	2.5 (0.1)	2.6 (0.1)	2.5 (0.1)	2.5 (0)	2.4 (0.1)	2.4 (0.1)	2.5 (0.1)	2.6 (0.1)	2.7 (0.1)	2.5 (0)	2.4 (0.2)	2.4 (0.1)
Saturated fats	3.6 (0.1)	3.5 (0.1)	3.5 (0.1)	3.4 (0.1)	3.6 (0.1)	3.6 (0.1)	3.6 (0.2)	4 (0.1)	3.4 (0.1)	3.5 (0.1)	3.5 (0.1)	3.5 (0)	3.6 (0.1)	3.3 (0.1)	3.5 (0.1)	3.7 (0.1)	3.8 (0.1)	3.5 (0)	3.8 (0.1)	3.5 (0.2)
Free sugars	6.2 (0.1)	6.3 (0.1)	6.1 (0.2)	6.2 (0.2)	5.5 (0.2)	6.4 (0.1)	5.2 (0.5)	6 (0.3)	6.3 (0.2)	6.1 (0.1)	6.6 (0.1)	6.6 (0.1)	5.6 (0.1)	5.7 (0.2)	6.9 (0.1)	6.2 (0.1)	6.2 (0.2)	6.1 (0.1)	7 (0.3)	7.5 (0.2)

Sodium	5 (0.1)	4.9 (0.1)	4.8 (0.1)	4.7 (0.2)	5.4 (0.2)	5 (0.1)	4.3 (0.4)	5.2 (0.3)	4.9 (0.2)	5 (0.1)	4.9 (0.1)	4.9 (0.1)	5.1 (0.1)	5 (0.1)	4.8 (0.1)	5.1 (0.1)	5.3 (0.2)	4.8 (0.1)	5.5 (0.3)	5.4 (0.3)
Total Score	39.4 (0.3)	42.3 (0.3)	40.5 (0.4)	39.4 (0.5)	39.3 (0.6)	41.8 (0.3)	38.9 (1)	42 (1)	40.7 (0.5)	40.6 (0.4)	41.1 (0.4)	41.8 (0.3)	39.1 (0.3)	36.9 (0.5)	42.1 (0.4)	42.3 (0.3)	42.9 (0.6)	40.2 (0.3)	41.4 (1)	41.9 (0.9)

Supplemental Table 2. Hazard ratios for gastrointestinal and aerodigestive cancer risk based on HEFI-2019 scores and covariates, Canadian Community Health Survey–Nutrition 2004 (CCHS 2004)–Canadian Cancer Registry linked cohort (N = 2,354,400)

Variable	Complete Case Analysis (N=2,354,400)
HEFI 2019 total score	1.01 (0.99, 1.04); p = 0.17
Sex	
Male	Reference
Female	1.22 (1.02, 1.47); p = 0.03
Educational attainment	
< High school grad	2.24 (1.52, 3.29); p < 0.001
High school grad	1.30 (0.85, 2.01); p = 0.22
Some post-secondary	0.82 (0.34, 1.97); p = 0.66
Post-secondary grad	Reference
Income	
Lowest income group	0.90 (0.37, 2.18); p = 0.82
Lower middle-income group	1.36 (0.66, 2.82); p = 0.41
Middle income group	1.77 (0.94, 3.34); p = 0.08
Upper middle-income group	1.13 (0.61, 2.10); p = 0.70
Highest income group	Reference
Marital status	
Married	Reference
Not married	1.15 (0.92, 1.45); p = 0.23
Smoking status	
Current smoker	Reference
Former smoker	1.69 (1.22, 2.34); p = 0.001
Never smoker	1.29 (0.97, 1.72); p = 0.08
Alcohol consumption	
No risk drinking	2.22 (1.45, 3.40); p < 0.001
Low risk	Reference
Moderate risk	1.21 (0.65, 2.24); p = 0.55
High risk	2.01 (1.08, 3.73); p = 0.03
BMI	1.02 (1.00, 1.04); p = 0.05
Age	1.03 (1.02, 1.04); p <0.001