Respiratory Health of Pacific Youth: nutrition resilience and risk factors in childhood

Siwei Zhai

Department of Statistics
The University of Auckland

Supervisor: Assoc. Prof. Alain C.Vandal, Dr Shabnam Jalili - Moghaddam

A thesis submitted in partial fulfilment of the requirements for the degree of MSc in Statistics, The University of Auckland, 2022.

Abstract

Background: The 2014 WHO Noncommunicable Diseases Country Profiles [1] shows that 7% of New Zealanders die from respiratory diseases. Particularly, the Pacific Island ethnic group is suffering deeply from these diseases. Compared to other ethnic groups, the Pacific Island ethnic group has higher hospitalization and mortality caused by respiratory diseases. The Pacific Islands Families (PIF) cohort study is an observational study which offers a reliable and unique data source to investigate the causal effect of risks and resilient behaviours/protective factors in early life on the lung function in early adulthood. The result of this study can help us access feasible solutions for public health intervention.

Objective: This study focused on finding the impact of modifiable nutritional risk and resilience factors in early life on the lung function in early adulthood for the Pacific Island ethnic group.

Methods: This study was based on data collected in Pacific Islands Families (PIF) cohort, focusing on respiratory health at age 18. We assigned variables collected during earlier assessments to 11 domains of risk and resilience factors. To make the result more interpretable, we applied three methods to further reduce the dimensions: combination of related variables into a single variable, selection of variables by subject-matter experts, and combination of variables into factor scores using factor analysis. Factor analysis was only implemented in the nutrition domain, which is the focus of this work. Before executing factor analysis, in order to make the food consumption comparable across measurement waves, we unified their unit to daily portions and classified them into 12 common food categories across measurement waves. In our study, Exploratory Factor Analysis (EFA) was employed to explore the underlying structure of food categories and to decide how many factors (eating patterns) were needed, prior to generating factor scores. In the following steps, we fitted a measurement invariance model, which is a multiple group model from Confirmatory Factor Analysis (CFA). This step aimed at estimating the loadings of food categories. This approach was selected to guarantee that the loadings were invariant across measurement waves and could be computed uniformly across any data set. Meanwhile, the factor scores based on these loadings should have the same invariant feature. We used weighted sum scores to compute the factor scores in this paper. This coarse method

upholds the invariance of the factor scores and reflect the impact of food category loadings on factors (eating patterns) in the factor scores. For causal inference, we used a causal diagram elicited from subject-matter experts to visualize the causal paths amongst the selected exposures, and implemented semi-parametric regression models (linear regression model and relative risk model) to obtain the causal results of nutrition factor scores on respiratory outcomes (FEV1 adjusted for height and sex, FEV1 Z-score, and FEV1 % predicted). It is worth noting that the response of relative risk model was the indicator based on the cutting point (-1.64) of healthy lung function in FEV1 Z-score. Since the PIFS cohort has suffered from attrition, possible selection bias needed addressing. To do so, we generated weights based on the baseline characters from the original birth cohort to reduce selection bias. In the last step, we computed population attributable fraction (PAF) of the nutrition factor scores to estimate the protected fraction of the healthy lung function due to nutrition at the population level, and also showed how the PAF is subject to the change of location of nutrition factor score in the particular eating pattern.

Results: In this study, we found that the eating patterns were basically align across all measurement waves with some differences. Amongst them, from the result of linear models, the ”Fruit and vegetables” eating pattern at 9 years had statistically and clinically significant significant causal effects on the healthy lung function in early adulthood. The higher nutrition factor scores in this eating patterns, the better lung function in the early adulthood. We estimated that, on average, one added portion per day of ”Fruit and vegetables” at 9 years will increase FEV1 Z-score by 0.25 units (95% CI: 0.00 - 0.43 units) or FEV1 % predicted by 2.94 percentage points (95% CI: 0.00 - 4.99 percentage points) or lung volume by 120 millilitre (95% CI: 0 - 210 millilitre). Furthermore, the PAF of healthy lung function showed the causal effect of the ”Fruit and vegetables” eating pattern at 9 years was also statistically significant at the population level. The results told that the consumption pattern of ”Fruit and vegetables” at 9 years is accountable for 11 percentage points of the prevalence of healthy lung function (95% CI: 0 - 19 percentage points), compared to o consumption at all. It offered a feasible way to enhance the healthy lung function prevalence amongst Pacific Island youth by a public health intervention - increases the average daily intakes of ”Fruit and vegetables” at 9 years.

limitations: 1. Nutrition factor scores used in the study may not be completely compatible with eating patterns from all measurement waves; 2. Some information was lost when unifying the food categories amongst all measurement waves in the study; 3. There may be some missing exposures in the causal diagram.

Strengths: 1. Nutrition factor scores were comparable over all measurement waves as their unit was unified; 2. As nutrition factor scores were expressed in daily portion, the PAF obtained in our study actually revealed how the prevalence of healthy lung function can be affected by a change in the number of daily portions of food consumption in the particular eating pattern; 3. The causal diagram was reviewed by experts in the relevant areas, so the generated results should be nearly unbiased; 4. Inverse probability weight (IPW) was used to guarantee a certain degree compensate for the impact of the selection bias.

Future research: We can 1. rerun the factor analysis on the 9-years measurement wave without following the early paper, and rebuild the models and recompute the PAF based on new nutrition factor scores; 2. use other methods to obtain the different weights; 3. examine various ways to change the distribution of the location of nutrition factor scores to reveal how the PAF of healthy lung function varies over different situations.

Contents

Abstract **1**

1 Introduction **7**
1.1 Pacific people in New Zealand . . . 7
1.2 Respiratory diseases and Pacific People . . . 7
1.3 Pacific Islands Families (PIF) cohort . . . 8
1.4 Respiratory health of Pacific youth study . . . 9
1.4.1 Respiratory risks in early life . . . 9
1.4.2 Resilience factors in early life . . . 10
1.4.3 Key outcome measure . . . 10
1.4.4 Aims for research . . . 11
1.4.5 Members of research team . . . 11

2 Method and theory **13**
2.1 Scope of this chapter . . . 13
2.2 Variable selection . . . 14
2.3 Factor analysis . . . 15
2.3.1 Food category mapping . . . 15
2.3.2 Factor determination . . . 27
2.3.3 Factor estimation . . . 28
2.4 Causal inference . . . 30
2.4.1 Causal graph . . . 30
2.4.2 Semi-parametric linear regression model . . . 31
2.5 Selection bias . . . 33
2.6 Population attributable fraction . . . 39

3 Data analysis **41**
3.1 Selected variables . . . 41
3.2 Dimensional reduction over Nutrition data . . . 55

3.2.1 Results from Exploratory Factor Analysis (EFA) 55
3.2.2 Results from Confirmatory Factor Analysis (CFA) 59
3.3 Causal diagram 67
3.4 Weight estimation 73
3.5 Causal results 80
3.5.1 Linear regression model 80
3.6 The results of population attributable fraction 86
3.6.1 Relative risk model 86
3.6.2 Population attributable fraction 87

4 Discussion 91
4.1 Summary of findings 91
4.2 Strengths and limitations 93
4.3 Future research 94

A Supplementary Figure 99
A.1 The distribution of inverse probability weights (IPW) 100

Chapter 1

Introduction

1.1 Pacific people in New Zealand

The Pacific people ethnic group is a collective concept including several ethnic groups from different Pacific Islands with diverse cultural backgrounds. In New Zealand this group is mainly comprised of Samoan, Cook Island Māori, Tongan, Niuean, Fijian, Tokelauan, Tuvaluan and Kiribati. Based on 2018 demographic information, Pacific people ethnic group accounted for 8.1% (381,642 people) of the New Zealand population (4,699,755 people), and 63.9% or 243,870 were living in Auckland, which is the largest city in New Zealand [2]. The age structure was youthful in this ethnic group. The median age was 23.4 years old and over half of people were under 25 years in 2018. Furthermore, the rate of growth of Pacific populations (10.8%) was much higher than the general population of New Zealand (29.0%) between 2014 and 2018 [2]. The Pacific population will contribute 10.7% to the population of New Zealand and 42.8% will be under 25 years within the ethnic group by 2043 [3]. This means that more people will be identified as Pacific people in New Zealand and this ethnic group will continue to have a young age structure in the future.

1.2 Respiratory diseases and Pacific People

Respiratory diseases form one of the most common category of diseases in New Zealand. The 2014 WHO Noncommunicable Diseases Country Profiles [1] shows that 7% New Zealanders die from respiratory diseases, making it the third deadliest group of diseases after cardiovascular diseases (32%) and cancer (29%). Meanwhile, we count approximately 69,000 hospitalisations per year due to respiratory diseases, with a rate of 1,563.1 per 100,000 people-years [4]. Respiratory diseases also correlate with the level of deprivation in an area. People living in the most deprived areas are more likely to be hospitalised or to die due to respiratory diseases compared to people living in the least deprived areas. Those in deprived areas have 2.9 times higher hospitalisations and 2.1 times greater mortality than those in least deprived areas [5]. The 2018 New Zealand census showed that 74% of the Pacific population lives in the country's most deprived areas [6]. Across all age groups, Pacific People in New Zealand have a 2.6 times higher hospitalisation rate for respiratory disease than other ethnic groups [4]. Identifying modifiable risk and resilience factors for respiratory diseases may help Pacific people enhance their health condition and provide support and orientation for the New Zealand health system to take further action.

1.3 Pacific Islands Families (PIF) cohort

The Pacific Islands Families (PIF) birth cohort study [7] was initiated by Dr Janis Paterson and Dr Colin Tukuitonga in 2000. An infant was considered a candidate (along with their family) if one of their parents identified him/herself as belonging to a Pacific Islands ethnicity and was a permanent resident of New Zealand. All candidates were recruited at Middlemore Hospital, South Auckland and had to be born between March and December 2000. The study within the PIF cohort could be broken down into several phases according to the crucial stages in the Pacific children's life [8, 9]: 1. First two years; 2. Transition to school; 3. Towards adolescence; 4. Early adulthood. There were multiple measurement waves within each phase. The measurement waves occurred at 6 weeks, 1 year and 2 years postpartum in the first phase; at 4 years and 6 years in the second phase; at 9 years, 11 years, and 14 years in the third phase; and at 18 years in the fourth phase. This study uses three major methods for data collection - maternal interview, paternal interview, and child assessment. It is worth noting that child assessments only started from the 4-year measurement wave. Furthermore, additional data from other sources was also collected, such as obstetric and perinatal information obtained from hospital records and postnatal information from Plunket. The PIF cohort has collected information regarding diverse exposures relevant to the health and development of children from the Pacific people ethnicity. Seemingly, it is a valid source for analysing how risk and resilience factors in early life may have an impact on respiratory health of Pacific people.

The modifiable risk and resilience factors in early life on later respiratory health were allocated them into 11 related domains for this paper - Immunisation, Exercise, Breastfeeding, Antenatal smoking, Smoking exposure, Smoking, Respiratory illness-infection, General health (Weight, height, BMI), Nutrition, Allergies, and Dwelling.

Table 1.1: Distribution of Observations based on domains and Measurement Wave

Domain	6 weeks	1 year	2 years	4 years	6 years	9 years	11 years	14 years
Immunisation	1,398	1,238	1,161	1,066	1,017		807	
Exercise				901			1,004	942
Breastfeeding	1,398		1,161					
Antenatal smoking	1,395							
Smoking exposure	1,398	1,236	1,154	1,060	1,018	1,004	1,047	952
Smoking		1,241						905
Respiratory illness-infection	1,397	1,240	1,162	1,054	1,019	1,013	1,047	951
Weight, height, BMI	1,379	1,231	1,041	1,066	1,018	890	1,002	918
Nutrition	1,398	1,241	1,162	907	801	976		204
Allergies				1,066	1,018	1,013	1,045	943
Dwelling	1,398	1,241	1,162	1,066	1,019	1,015	1,043	897

1.4 Respiratory health of Pacific youth study

1.4.1 Respiratory risks in early life

Modern research indicates that adult respiratory health could be causally related to early events occurring from infancy to childhood [10, 11]. This is a crucial period for the development of lungs, which is the main organ in the respiratory system. The development of lungs starts in utero and continues in early childhood. Lung function peaks at around 20-22 years-old for male and around 18-20 years for female, and then decreases with age. Good lung function means a strong respiratory system and less chance of contracting respiratory diseases. However, some early events could weaken lung function/respiratory system in two ways: 1) Lowering the peak of lung function and/or speeding up the decline in lung function after the peak; 2) Inducing early respiratory diseases and increasing susceptibility to developing a later disease [12]. If modifiable risk could be identified from early events and removed or changed in child's life, it could be an effective way to enhance adult respiratory health. The PIF Respiratory health of Pacific youth study considered factors from early life (birthweight, antenatal smoke exposure, postnatal smoke exposure, allergies, dwelling conditions from the first 2 years of life) and childhood (smoking at 14 years), as the main modifiable risk in early life [13].

1.4.2 Resilience factors in early life

Respiratory function and adult respiratory disease may be also affected by resilience or protective factors, such as resilience or protective factors, such as breastfeeding, immunisation, and conceivably, physical activity levels and nutrition [14, 15]. However, few studies have examined the topic of childhood resilience factors promoting later respiratory health. Early pathological studies had concluded that alveolar structure completely forms at two years of age, but more recent magnetic resonance imaging reveals that alveoli continue growing after that and fully develop in early adulthood [16–18]. This finding illustrates that resilience factors may help later lung growth at a certain period. The major resilience factors considered the PIF Respiratory health of Pacific youth study are factors from early life (breastfeeding, immunisation and nutrition from the first 2 years of life) and childhood (nutrition and exercise at 4-years, 11-years and 14-years of age) [13].

1.4.3 Key outcome measure

Forced expiratory volume in 1 second (FEV1) assessed by spirometer is the key output measure in this research. It is a significant marker for measuring the level of respiratory health. FEV1 can be used to screen, diagnose, and monitor respiratory diseases such as asthma, brochiectasis, cystic fibrosis, and chronic obstructive pulmonary disease (COPD) [19–22]. It is also an important indicator to measure the progress of lung disease, lung transplantation referral [23, 24] and death [23, 25]. ”FEV1 is more than a measure of airflow limitation, but a marker of premature death with broad utility in assessing baseline risk of chronic obstructive pulmonary disease (COPD), lung cancer, coronary artery disease and stroke” (Young, Hopkins, & Eaton, 2007) [26]. We will use the variables related to FEV1 (FEV1 adjusted for height and sex, FEV1 Z-score, and FEV1 % predicted) as the response in the causal models to measure the impact of modifiable risk or the resilience factors in childhood on respiratory health in adulthood. FEV1 Z-score is the primary outcome measure while FEV1 adjusted for height and sex and FEV1 % predicted are used to better interpret the result. Particularly, FEV1 adjusted for height and sex can be used to identify whether the result is clinical significant since a change of 100 mL in FEV1 will significantly impact lung function [27]. Specifically, in this thesis we focus on the impact of nutrition factors in childhood on respiratory health in adulthood.

1.4.4 Aims for research

The PIFS Respiratory Health of Pacific Youth Study intended to achieve three aims (El-Shadan, Conroy, Alain, Shabnam, Emily, Leon, Adrian, & Catherine, 2020) [13]:

> (i) Estimate the effect of early life (eg, birthweight, antenatal smoke exposure, postnatal smoke exposure) and childhood risk factors (eg, allergies, dwelling conditions from the first 2 years of life, child smoking at 14 years) on peak lung function attainment and respiratory outcomes in Pacific youth aged 18 to 19 years;
>
> (ii) Determine modifiable childhood risk and protective factors; including breastfeeding, immunization, and nutrition during the first 2 years of life; exercise at ages 4, 11, and 14 years; peak flow at ages 6 and 9 years; respiratory infections, respiratory condition–related hospital admissions, and reported breathing problems in the first 2 years of life; and asthma in childhood) on lung function attainment and respiratory outcomes in Pacific youth aged 18 to 19 years;
>
> (iii) Estimate the population attributable fraction and population avoidable fraction of modifiable early life risk factors and childhood resilience factors on these outcomes.

The present thesis focuses on childhood nutritional risk and protective factors.

1.4.5 Members of research team

The PIF Respiratory health of Pacific youth study investigative team consists of the following members:
Name: Dr El-Shadan Tautolo
Department: Pacific Health Research Centre
Organisation: Auckland University of Technology
Role in project: Principal Investigator

Name: Associate Professor Catherine Byrnes
Department: Department of Paediatrics
Organisation: University of Auckland
Role in project: Clinical Leadership and support

Name: Dr Conroy Wong
Department: Respiratory Medicine
Organisation: Counties Manukau District Health Board
Role in project: Clinical Research Support

Name: Associate Professor Alain Vandal
Department: Department of Statistics
Organisation: University of Auckland
Role in project: Senior Biostatistician

Name: Leon Iusitini
Department: Pacific Health Research Centre
Organisation: Auckland University of Technology
Role in project: Study Coordinator

Name: Dr Shabnam Jalili - Moghaddam
Department: National Institute Stroke & Applied Neurosciences
Organisation: Auckland University of Technology
Role in project: Study Coordinator

Chapter 2

Method and theory

2.1 Scope of this chapter

Although this thesis focuses on nutrition as an exposure, we provide some detail concerning other exposure domains to indicate how we identified and distinguished the nutrition domain from these other domains.

To obtain the estimated causal effects of the modifiable risk factors and resilience factors (hereafter, ”exposures”) on the respiratory outcomes, we faced several issues that needed to be solved:

- How to select the most related variables out of the several thousand variables collected thus far from the PIFS cohort. (Section: 2.2)
- How to reduce the dimensionality of some of the exposure variables involved, so as to capture the constructs involved validly while maintaining modifiablility.(Section: 2.3)
- How to set up the causal diagram, to identify confounders and mediators (and any eventual collider) of exposures on the respiratory outcomes. (Section: 2.4.1)
- How to select a suitable statistical model to model the causal relationship between exposures and the respiratory outcomes, and estimate the causal effects. (Section: 2.4.2)
- How to handle the missingness and attrition in the PIFS cohort. (Section: 2.5)
- What proportion of the respiratory outcomes is attributable to the effect of a particular risk factor or a particular modifiable protective factor at the population level. (Section: 2.6)

The following sections will introduce the solutions used in our research to address these issues.

2.2 Variable selection

Information on more than 10,000 variables has been collected in the PIF cohort so far, and around 1,300 variables can be considered as risk or protective factors for respiratory health. Their frequencies can be shown by domains and measurement waves.

Table 2.1: Distribution of Variables based on domains and Measurement Wave

Domain	**6 weeks**	**1 year**	**2 years**	**4 years**	**6 years**	**9 years**	**11 years**	**14 years**
Immunisation	1	5	7	4	5		1	
Exercise				9			2	34
Breastfeeding	49		4					
Antenatal smoking	22							
Smoking exposure	2	3	6	2	10	5	17	14
Smoking		6						3
Respiratory illness-infection	2	3	4	2	16	8	4	1
Weight, height, BMI	9	2	2	5	16	10	12	32
Nutrition	35	33	49	210	222	37		83
Allergies				1	2	8	5	2
Dwelling	71	63	82	14	60	30	31	22

As Table 2.1 shows, some of the domains have a large number of variables at some particular measurement waves. Examples are the Breastfeeding domain at 6-weeks measurement wave; the Nutrition domain at 4-years and 6-years measurement waves; and the Dwelling domain at most of the measurement waves. Although it is conceivable to estimate the effect of every variable on respiratory outcome, we elected to represent the domains with a small number of exposures at each measurement wave in the hope to simplify interpretation of the results. For this purpose, we reduced the number of variables in each domain by:

- Combining related variables as a single variable interpretably, such as combining the number of days vigorous physical activities, usual hours/day vigorous activities, and usual minutes/day vigorous activities as total hours per week vigorous activities.
- Selecting variables according to expert opinion. Associate Professor Catherine Byrnes and Dr Conroy Wong are respiratory physicians in the research team. They utilised their professional knowledge to select representative and clinically meaningful exposure variables in each domain, with the exception of the Nutrition domain.
- Produce weighted averages of variables using factor analysis. This method was only

implemented on the nutrition domain. In 2018, Dr Shabnam Jalili - Moghaddam applied factor analysis on nutrition variables at 14-years measurement wave and identified several nutrition domains [28]. This thesis follows and extends her approach, under her guidance, to nutrition variables at 4-years, 6-years, and 9-years measurement wave.

- Filtering out variables with less completeness.

2.3 Factor analysis

2.3.1 Food category mapping

Researchers did not use an identical dietary assessment method at all measurement waves. The 4-years and 6-years measurement wave shared the same Food Frequency Questionnaire (FFQ). This questionnaire surveyed the consumption of single food items, such as banana, peach, and burger, by asking how often they were consumed. By contrast, the dietrary habit questionnaire was applied at 9-years and 14-years measurement wave inquired about the consumption for a food group such as fruit, vegetable, and red meat, and required answers in intake frequency or portions per day. In order to make the food consumption comparable among measurement waves, we defined 12 food categories common to all measurement waves, and mapped single food items at 4-years and 6-years and food groups at 9-years and 14-years to related food category. These food categories were defined according to Dr. Shabnam Jalili - Moghaddam's previous work [28] and further guidance. Compared to her original work, there are three fewer food categories in this thesis. Hot chips, French fries, wedges/ kumara chips and Battered/ fried fish/ shellfish are merged into Fast food/ takeaways. Milk is removed as it is completely missing at 9-years. The detail of the mappings of questionnaire items to food categories is presented in Table 2.2. The original variable names are presented in the table to facilitate reproducibility.

Table 2.2: The mappings between variables and food categories

Food category	Question at 4-years and 6-years	Variable at 4-years	Variable at 6-years	Question at 9-years and 14-years	Variable at 9-years	Variable at 14-years
Fast food/ takeaways	How often does your child eat fried chicken or chicken nuggets	t4f_f49a	t6f_f49a	How often do you eat fast food or takeaways from places like McDonalds, Burger King, Pizza shops, or fish and chips shops?	t9pe25	t14ag19
	How often does your child eat fried fish or takeaway fish or raw fish with coconut cream or milk	t4f_f50a	t6f_f50a	How often do you eat battered or fried fish or shellfish?	t9pe14	t14ag12
	Fish cake, fish fingers or fish pie	t4f_oft51	t6f_oft51	How often does your child eat hot chips, French fries, wedges, or kumara chips? Think about lunch and dinner as well as snacks.	t9pe24	t14ag18
	Meat pie	t4f_oft56	t6f_oft56			

Continued on next page

Table 2.2 – *Continued from previous page*

Food category	**Question at 4-years and 6-years**	**Variable at 4-years**	**Variable at 6-years**	**Question at 9-years and 14-years**	**Variable at 9-years**	**Variable at 14-years**
	Burgers	t4f_oft57	t6f_oft57			
	Sausage rolls	t4f_oft60	t6f_oft60			
	Pizza	t4f_oft74	t6f_oft74			
	Potato crisps, corn snacks or chips	t4f_oft97	t6f_oft97			
	Popcorn	t4f_oft98	t6f_oft98			
	Fried potatoes	t4f_oft11	t6f_oft11			
	Cooked green banana mainly with coconut cream	t4f_oft17	t6f_oft17			
Soft drinks/ energy drinks	Coca cola or other cola drinks	t4f_oft109	t6f_oft109	How often do you drink soft drinks or energy drinks?	t9pe27	t14ag21
	Mountain Dew	t4f_oft110	t6f_oft110			
	New Age' drinks, eg. V.E, Red Bull	t4f_oft111	t6f_oft111			
	Soft drinks, eg. lemonade, orange	t4f_oft112	t6f_oft112			
	Sport drinks, eg. Gatorade, Powerade	t4f_oft113	t6f_oft113			

Continued on next page

Table 2.2 – *Continued from previous page*

Food category	**Question at 4-years and 6-years**	**Variable at 4-years**	**Variable at 6-years**	**Question at 9-years and 14-years**	**Variable at 9-years**	**Variable at 14-years**
Lollies, sweets, chocolate and confectionary	Chocolate coated or cream filled biscuits	t4f_oft85	t6f_oft85	How often do you eat lollies, sweets, chocolate, and confectionary?	t9pe28	t14ag22
	Biscuits	t4f_oft86	t6f_oft86			
	Bars	t4f_oft87	t6f_oft87			
	Crackers or crispbreads	t4f_oft88	t6f_oft88			
	Cake or slice	t4f_oft89	t6f_oft89			
	Doughnuts rings (deep fried) or croissants	t4f_oft90	t6f_oft90			
	Pancake or pikelets	t4f_oft92	t6f_oft92			
	Fruit pie, fruit crumble or tart	t4f_oft93	t6f_oft93			
	Pudding	t4f_oft94	t6f_oft94			
	Custard or custard puddings	t4f_oft95	t6f_oft95			
	Chocolate	t4f_oft99	t6f_oft99			
	Candy coated chocolate	t4f_oft100	t6f_oft100			

Continued on next page

Table 2.2 – *Continued from previous page*

Food category	**Question at 4-years and 6-years**	**Variable at 4-years**	**Variable at 6-years**	**Question at 9-years and 14-years**	**Variable at 9-years**	**Variable at 14-years**
Fruit juices and fruit drinks	Juice	t4f_oft106	t6f_oft106	How often do you drink fruit juices and fruit drinks?	t9pe26	t14ag20
	Powdered fruit drink, eg. Refresh, Raro	t4f_oft107	t6f_oft107			
	Fruit drink from concentrate or cordial, eg. Just Juice, Ribena	t4f_oft108	t6f_oft108			
Canned fish/ shellfish	Canned fish, eg. Tuna or salmon	t4f_oft52	t6f_oft52	How often do you eat canned fish or shellfish?	t9pe15	t14ag13
Fresh/ frozen fish/ shellfish	Fish	t4f_oft50	t6f_oft50	How often do you eat fresh or frozen fish or shellfish?	t9pe13	t14ag11
	Shell fish, eg. mussel, paua or crabmeat includes lobster	t4f_oft53	t6f_oft53			

Continued on next page

Table 2.2 – *Continued from previous page*

Food category	**Question at 4-years and 6-years**	**Variable at 4-years**	**Variable at 6-years**	**Question at 9-years and 14-years**	**Variable at 9-years**	**Variable at 14-years**
Vegetables (Fresh/ frozen/ canned)	Other potatoes	t4f_oft12	t6f_oft12	On average how many servings of vegetables - fresh, frozen or canned - do you eat per day?	t9pe17	t14ag15
	Taro mainly with coconut cream	t4f_oft13	t6f_oft13			
	Kumara	t4f_oft14	t6f_oft14			
	Carrots	t4f_oft15	t6f_oft15			
	Cassava	t4f_oft16	t6f_oft16			
	Pumpkin	t4f_oft18	t6f_oft18			
	Mixed vegetables	t4f_oft19	t6f_oft19			
	Corn	t4f_oft20	t6f_oft20			
	Peas	t4f_oft21	t6f_oft21			
	Silverbeet, spinach, taro leaves, puha or watercress	t4f_oft22	t6f_oft22			
	Green beans	t4f_oft23	t6f_oft23			
	Broccoli	t4f_oft24	t6f_oft24			
	Cauliflower or cabbage	t4f_oft25	t6f_oft25			

Continued on next page

Table 2.2 – *Continued from previous page*

Food category	**Question at 4-years and 6-years**	**Variable at 4-years**	**Variable at 6-years**	**Question at 9-years and 14-years**	**Variable at 9-years**	**Variable at 14-years**
	Roast vegetables	t4f_oft26a	t6f_oft26a			
	Lettuce or green salad	t4f_oft27	t6f_oft27			
	Tomatoes	t4f_oft28	t6f_oft28			
	Capsicum	t4f_oft29	t6f_oft29			
	Avocado	t4f_oft30	t6f_oft30			
Fruit (fresh/ frozen/ canned/ stewed)	Banana, raw	t4f_oft1	t6f_oft1	On average how many servings of fruit - fresh, frozen, canned or stewed - do you eat per day?	t9pe16	t14ag14
	Apples or pears	t4f_oft2	t6f_oft2			
	Oranges or mandarins	t4f_oft3	t6f_oft3			
	Kiwifruit	t4f_oft4	t6f_oft4			
	Nectarines, peaches, plums or apricots	t4f_oft5	t6f_oft5			
	Strawberries or other berries	t4f_oft6	t6f_oft6			

Continued on next page

Table 2.2 – *Continued from previous page*

Food category	**Question at 4-years and 6-years**	**Variable at 4-years**	**Variable at 6-years**	**Question at 9-years and 14-years**	**Variable at 9-years**	**Variable at 14-years**
	Canned or cooked fruit	t4f_oft7	t6f_oft7			
	Dried fruit	t4f_oft8	t6f_oft8			
Red meat	Roast beef, lamb or pork	t4f_oft40	t6f_oft40	How often do you eat red meat?	t9pe8	t14ag5
	Steak	t4f_oft41	t6f_oft41			
	Lamb or mutton chops	t4f_oft42	t6f_oft42			
	Pork chop (or other pork small cuts)	t4f_oft43	t6f_oft43			
	Boiled corned beef/ silverside includes brisket	t4f_oft44	t6f_oft44			
	Mince, including rissoles, patties, Shepherd's Pie,etc	t4f_oft46	t6f_oft46			
	Liver or liver pate	t4f_oft47	t6f_oft47			
Chicken	Chicken	t4f_oft49	t6f_oft49	How often do you eat chicken?	t9pe9	t14ag7
Processed meat products	Canned corned beef mainly full fat or low fat	t4f_oft45	t6f_of45fat	How often do you eat processed meat products?	t9pe12	t14ag9

Continued on next page

Table 2.2 – *Continued from previous page*

Food category	**Question at 4-years and 6-years**	**Variable at 4-years**	**Variable at 6-years**	**Question at 9-years and 14-years**	**Variable at 9-years**	**Variable at 14-years**
	Bacon or ham	t4f_oft48	t6f_oft48			
	Sausages	t4f_oft58	t6f_oft58			
	Luncheon, ham and chicken	t4f_oft59	t6f_oft59			
Bread, including toast and bread rolls	Bread, including toast and bread rolls	t4f_oft62	t6f_oft62	On average, how many slices of bread/toast OR bread rolls do you eat per day?	t9pe2	t14ag2

However, some food items at 4-years and 6-years are not able to be assigned to any of these food categories. We therefore decided to exclude them from the study. Table 2.3 gives the list of excluded food items.

Table 2.3: List of excluded variables at 4-years and 6-years in Nutrition domain

Question	**Variable at 4-years**	**Variable at 6-years**
Eggs, boiled, poached, fried or scrambled, etc.	t4f_oft39	t6f_oft39
Meat and vegetable 'boil-up'	t4f_oft33	t6f_oft33
Meat stew or casserole with vegetables	t4f_oft34	t6f_oft34
Pasta with meat and tomato sauce	t4f_oft35	t6f_oft35
Pasta with cream, white sauce or cheese sauce	t4f_oft36	t6f_oft36
Chinese type dishes, stir-fry meat or chicken and vegetables includes chop suey	t4f_oft37	t4f_oft37
Breakfast cereal	t4f_oft63	t6f_oft63
Rice	t4f_oft64	t6f_oft64
Fried rice		t6f_oft64b
Jam or honey	t4f_oft66	t6f_oft66
Nutella	t4f_oft67	t6f_oft67
Marmite or Vegemite	t4f_oft68	t6f_oft68
Peanut butter	t4f_oft69	t6f_oft69
Mayonnaise or salad dressing, including coconut cream	t4f_oft70	t6f_oft70
Tomato sauce or ketchup	t4f_oft71	t6f_oft71
Gravy	t4f_oft72	t6f_oft72
Soup	t4f_oft75	t6f_oft75
Noodles	t4f_oft76	t6f_oft76
Canned spaghetti with tomato sauce	t4f_oft77	t6f_oft77
Baked beans	t4f_oft78	t6f_oft78

Continued on next page

Table 2.3 – *Continued from previous page*

Question	**Variable at 4-years**	**Variable at 6-years**
Ice cream	t4f_oft80	t6f_oft80
Cheese	t4f_oft81	t6f_oft81
Yoghurt or Dairy food	t4f_oft82	t6f_oft82
Cream	t4f_oft83	t6f_oft83
Ice blocks	t4f_oft114	t6f_oft114
Tea	t4f_oft115	t6f_oft115
Coffee	t4f_oft116	t6f_oft116
Milk (not flavoured)	t4f_oft102	t6f_oft102
Flavoured milk	t4f_oft103	t6f_oft103
Milk shake	t4f_oft104	t6f_oft104
Food dirnk, eg. Milo powder, Nesquik	t4f_oft105	t6f_oft105
Butter or margarine vegetables	t4f_oft26b	t6f_oft26b

As indicated earlier, the available answers regarding food consumption were not identical at different measurement waves. In the 4-years' and 6-years' questionnaire, available answers consisted in discrete choices in terms of intake frequency; while in the 9-years' and 14 years' questionnaire answers were in terms of intake frequency or daily intake.

We converted all responses to daily intake in this paper. Table 2.4, 2.5, and 2.6 provide the details on the conversion criteria. We then obtained the total consumption of a food category by summing up the consumption of food items or food groups mapped to this category. There are 4 food consumption in each food category, corresponding to the 4 measurement waves. One point to note is that any values over the upper bound of the original scale of the food category were reduced to its the upper bound across all measurement waves. This guarantees that the food consumption of the same food category is located in the same range, across all ages.

Table 2.4: Frequency of consumption of food and the weighting factor applied so as to standardise to a daily rate (4-years, 6-years, 9-years, and 14-years)

Frequency of consumption	**Weighting factor /day**
Never or less than once a month	1/200
1-3 times a month	2/30
1-2 times a week	1.5/7
3-4 times a week	3.5/7
5-6 times a week	5.5/7
Once a day	1
More than once a day	2

Table 2.5: Frequency of consumption of food and the weighting factor applied so as to standardise to a daily rate (9-years and 14-years)

Frequency of consumption	**Weighting factor /day**
None	0
Less than one per day	0.5
1-2 per day	1.5
3-4 per day	3.5
5-6 per day	5.5
7 or more per day	7

Table 2.6: Number of servings of consumption of food per day and the applied weighting factor so as to standardise to a daily rate (9-years and-14 years)

Number of servings per day	**Weighting factor /day**
Never	0
Less than one serving per day	0.5
1 serving	1
2 servings	2
3 servings	3
4 or more servings	4

2.3.2 Factor determination

There are approximately 600 variables in the nutrition domain. Since we were interested in a global picture of the effect of nutrition exposure on respiratory health in young adulthood, rather than the effect of specific foodstuffs, we undertook dimensional reduction before the analysis. Initial attempts at using sliced inverse regression [29] and other outcome-dependent dimensional reduction approaches [30] did not yield interpretable nutrition summaries. Instead, we to previously validated work on nutrition in the PIF cohort [28] and turned to factor analysis.

Briefly, all nutrition items transformed to daily portion of food categories at all measurement waves were entered in a multi-group confirmatory factor analysis (CFA) model assuming measurement invariance across the measurement waves, and 4 nutrition domains, adapted to the available data, as identified in [28]. Factor scores were computed as weighted average of the items involved, with weights being the loadings common to all measurement waves (method 4 in [31]). Exploratory factor analysis (EFA) assuming four factors was also carried out at each wave and the results compared with those appearing in [28], to assess the quality of the nutrition summaries.

Exploratory Factor Analysis (EFA) is a factor analysis method to investigate the underlying structure in the observed measures. It is a data-driven approach as there is not any priori conditions assumed in the exploration process [32]. This technique loads latent factors to sets of correlated variables in an attempt to account for the correlations, and these factors can be used to summarise the information from each set of observed measures. EFA can thus be used to find a small set of appropriated latent factors to represent a large set of observed measures, allowing a reduced dimensions to be used in later analyses. Original work on nutrition in the PIF cohort [28] yielded four latent factors (eating patterns) from the measurement wave at 14-years by using EFA. Compared to this earlier study, the current thesis covers three more measurement waves (4-years, 6-years, and 9-years) and applies some adjustments to the food categories and the related mappings. Rerunning EFA with the same factors based on the redefined food categories helped us reassess and adjust the original eating patterns to accustom the change we made in this thesis. The rebuilt results were generated with the use of psych package [33] in R, and its default rotation method for EFA is the oblimin transformation. We then compared the original eating patterns (based on 14-years measurement wave) to the rebuilt eating patterns (based on 4-years, 6-years, 9-years, and 14-years measurement wave respectively) to assess the difference between them. Aside from the application of data-driven numerical, we strove to remain aligned with the previously published results, which were strongly inferred by substantive aspects of nutrition science. Therefore, in this thesis, we will follow the original eating patterns.

2.3.3 Factor estimation

We aimed to obtain factor scores for the eating patterns and used them to represent the information from the nutrition domain in subsequent analysis. Factor scores are numerical representation of the latent variables identified through factor analysis, and there are several approaches to their computation. Selection of the approach was informed by the need for factor scores to remain invariant across measurement waves, remain calculable with any new data set and be clearly interpretable. This condition ensures that nutrition factor scores are identifiable, measurable and modifiable in any new population, allowing for resppiratory health-based public health interventions on nutrition to be well-defined in terms of impact (assuming there is any causal relation between early childhood and adolescent nutrition and respiratory health). For these reasons, we elected to compute factor scores as weighted average of the daily portions. The weights were taken to be the standardized loadings, corresponding to method 4 of [31]. Since the factor scores are derived from the factor loadings, the first step is to calculate the factor loadings and they will have the same requirement of invariance as the factor scores. In this thesis, we employed confirmatory factor analysis (CFA) to obtain the factor loadings. Compared to EFA, CFA will generate the factor loadings according to the known factor solution. Therefore, it is necessary to prespecify the number of factors, the factor structure, and the constraints in the factor structure. In general, they are given by theoretical and substantive aspects of subject matter [33]. This thesis used the eating patterns defined in the previous substantive research [28] as the known priori factor structure and assumed that this factor structure was suitable for all measurement waves. Hence, we implemented the multiple group modeling technique on CFA to impose the same factor structure on all groups. Its mathematical model is [34, 35]:

$$y_i^g = \tau_i^g + \Lambda_{y_{ij}}^g \eta_j^g + \epsilon_i^g,\ i = 1, \ldots, q;\ j = 1, \ldots, m;\ g = 1, \ldots, G$$

where

- y_i^g is the i^{th} observed measure in the g^{th} group
- τ_i^g is the intercept of the i^{th} observed measure in the g^{th} group.
- $\Lambda_{y_{ij}}^g$ is the loading of the i^{th} observed measure on the j^{th} latent factor in the g^{th} group
- η_j^g is the j^{th} latent factor in the g^{th} group.
- ϵ_i^g is the random measurement error of the i^{th} observed measure in the g^{th} group

Furthermore, the latent factors can be further explained by the following equation:

$$\eta_j^g = \alpha_j^g + \zeta_j^g,\ j = 1, \ldots, m;\ g = 1, \ldots, G$$

where

- η^g_j is the j^{th} latent factor in the g^{th} group.
- α^g_j is the mean of the j^{th} latent factor in the g^{th} group.
- ζ^g_j is the residual of the j^{th} latent factor in the g^{th} group.

and the covariance formula to link the measurement structure to the manifest covariance matrix is:

$$\Sigma^g(\Theta) = \Lambda^g_{y_{ij}} \Psi^g_j (\Lambda^g_{y_{ij}})' + \Theta^g_{\epsilon_{y^g_{ij}}}$$

where

- $\Sigma^g(\Theta)$ is the covariance matrix of the i^{th} observed measure in the g^{th} group
- Ψ^g_j is the covariance matrix of the j^{th} latent factor.
- $(\Lambda^g_{y_{ij}})'$ is the transpose of $\Lambda^g_{y_{ij}}$
- $\Theta^g_{\epsilon_{y^g_{ij}}}$ is measurement error variances of manifest variables y^g_{ij}

We fitted two CFA multiple group models in this thesis. They are the configural invariance model and the structural invariance model. The former is the basic model without any constraints on the loading (This can be replaced by the EFA models) while the latter is the model with strong constraints to fulfil the invariance requirement. The factor loadings ($\Lambda^A = \Lambda^B = ... = \Lambda^G$), variances ($\Psi^A_{jj} = \Psi^B_{jj} = ... = \Psi^G_{jj}$), and covariances ($\Psi^A_{jk} = \Psi^B_{jk} = ... = \Psi^G_{jk}$) are constrained so as to be equal across groups in the structural invariance model. The configural invariance model is used as the control model to evaluate the performance of the structural invariance model and assisted us to remove the ineffective food categories from the known priori factor structure based on R square. However, regardless of the result of the evaluation, we will adhere to using the factor loadings estimated by the structural invariance model in the factor score calculation. This is because our purpose was not to produce a well-fitting CFA model, but rather a weighting scheme with reasonable face validity informed by the previous substantive research. We employed lavaan package [36] to build these models in R. R-squared ($R^2 = 1 - \frac{RSS}{TSS}$ where R^2 = coefficient of determination, RSS = sum of squares of residuals, TSS = total sum of squares) is a statistical technique to reveal how well a regression model to explain the variability of an exposure. In this thesis, we applied R-squared to compare the performance of different invariance models on the explanation of the variance of the food categories, and determined whether an exposure will be kept in the CFA model based on its R-squared.

Weighted sum scores [31] is the method used to estimate the factor scores in this paper. It is a non-refined method so does not involve sophisticated technical computation. In this method, the factor scores are only determined by the consumption of food categories and the factor loadings from the structural invariance model. This can guarantee that the invariance requirement is not changed by the computation. The weighted sum scores method generates the factor scores by utilising the following steps:

1. Obtain the weights in each nutrition domain as the proportion of each food category loading to the sum of the loadings in the corresponding factor.

2. Rescaling the consumption of the food categories by multiplying the weighted loadings.

3. Generating the factor scores by summing up the scaled consumption of the food categories within the factor.

We applied this method on every factor in each measurement wave to obtain a full set of factor scores. In this thesis, the factor score can be interpreted as a weighted average of portions per day for an eating pattern, therefore has portions per day as an unit. Within the weighted sum scores method, the food category with the higher loadings will have the larger impact on the factor scores. This means that the impact of the loadings on the factor can be revealed in the factor scores. However, there are some drawbacks to this method as well. The factor scores will be correlated if there are correlation in the fitted loading pattern. This is because the non-refined method does not have any extra computation to correct the correlation, and this correlation may differ from the correlation between the factors. This is because non-refined methods do not involve any extra computation to correct the correlation.

2.4 Causal inference

2.4.1 Causal graph

The first step in causal inference is to obtain a set of causal assumptions between exposures, X, and response, Y. It is a non-mathematical process. We settled on these assumptions based on the suggestions from domain experts, and utilised causal graph [37] as a tool to visualize them. It offers a better way to understand and infer the causal relationship between X and Y. Particularly, confounding paths between X and Y can be clearly represented in a causal graph. This was crucial in assisting us in building the causal model in the later step. Confounders, Z, will interfere with the estimate of the causal effects of exposures on the outcome as they covary with X and can independently generate an impact on Y. To reduce/remove the confounding effects between X and Y, Z shall be included by the causal model so as to block confounding paths. In this thesis, we used R package dagitty [38] and ggdag [39] package to generate the causal graph and identify adjustment sets. We follow two basic assumptions when drawing this graph: 1. All variables have direct impact on the final outcome; 2. Only the variable which came from the previous measurement wave can affect the variable in the current measurement wave, if there is a causal relationship between them.

2.4.2 Semi-parametric linear regression model

The parametric regression model is the usual way to estimate the causal effects if we an make an assumption regarding the underlying error distribution from prior knowledge. It is easier to understand and interpret than other types of models. However, the standard error estimates may be biased when the assumption is wrong. In fact, it is hard to obtain enough prior knowledge to correctly assume the true error distribution of a particular dataset in the real world . A better way is to obtain the required estimates without these assumptions. We decided to utilize semi-parametric regression models to achieve this in our research. It is a hybrid model composed of two parts: parametric and non-parametric. The parametric part will give the estimate of the causal effect of the exposure on the outcome conditional on the confounders by solving estimating equations based on the form of the linear or generalised linear model. It can guarantee consistent estimates, regardless of the true error distribution. Meanwhile, the non-parametric part will be used to estimate the error distribution. In this thesis, we used bootstrap method to effect this. This method does not rely on the error distribution being correctly specified, but the empirical error distribution in the dataset. It will determine a reasonable estimate for the true error distribution from a large number of simulations based on sampling with replacement.

In this thesis, we fitted two different semi-parametric regression models - linear regression model and relative risk model. Linear regression model enable us to interpret the average causal effect of a nutrition factor score on the mean of respiratory outcomes (1. FEV1 adjusted for height and sex: Forced expiratory volume in 1 second (FEV1) adjusted for child's height and sex; 2. FEV1 Z-score: Forced expiratory volume in 1 second (FEV1) adjusted by standard deviation (SD) ; 3. FEV1 % predicted: Forced expiratory volume in 1 second (FEV1)/Forced vital capacity (FVC) ratio of the patient divided by the average FEV1/FVC ratio in the population for any person of similar age, sex, and body composition) in terms of the daily portion of a particular eating pattern. The equation of our linear regression model is:

$$E[y_i|x_i, z_{1i}, ..., z_{ki}] = \alpha + \beta x_i + \sum_{j=1}^{k} \gamma_{ji} z_{ji} + \varepsilon_i$$

where

- $E[y_i|x_i, z_{1i}, ..., z_{ki}]$ is the expectation of the i^{th} response conditional on the i^{th} exposure and its related confounders.
- y_i is the i^{th} response (FEV1 adjusted for height and sex, FEV1 Z-score, or FEV1 % predicted).
- α is the intercept.
- β is the coefficient of the i^{th} exposure.

- x_i is the i^{th} exposure.
- γ_{ji} is the coefficient of the j^{th} confounder of the i^{th} exposure.
- z_{ji} is the j^{th} confounder of the i^{th} exposure.
- ε_i is the residual of the expectation of the i^{th} response conditional on the i^{th} exposure and its related confounders.

The other model, relative risk model, offers us a ratio of risks between two different groups, and assist us to obtain the population attributable fraction (PAF). In this research, we split the data set into two groups based on the Lower Limit of Normal (LLN = -1.64) of FEV1 Z-score - Group 1: FEV1 Z-score >= -1.64; Group 2: FEV1 Z-score $<$ -1.64 [40]. Typically, FEV1 Z-score $<$ -1.64 indicates unhealthy lung function [41]. As we plans to estimate the beneficial causal effect of nutrition factor scores on the lung function, Group 2 is considered as the reference group. To obtain the ratio of risks between these two groups, we created an indicator based on these labels (Healthy lung function indicator) and set it as the response for relative risk model. Based on this setting, we can obtain the ratio of risks. The equation of our relative risk model is:

$$ln(E[p_i = 1|x_i, z_{1i}, ..., z_{ki}]) = \alpha + \beta x_i + \sum_{j=1}^{k} \gamma_{ji} z_{ji} + \varepsilon_i$$

where

- $E[p_i = 1|x_i, z_{1i}, ..., z_{ki}]$ is the expected probability that the i^{th} response is equal to 1 conditional on the i^{th} exposure and its related confounders.
- y_i is the i^{th} response (Healthy lung function indicator).
- α is the intercept.
- β is the coefficient of the i^{th} exposure.
- x_i is the i^{th} exposure.
- γ_{ji} is the coefficient of the j^{th} confounder of the i^{th} exposure.
- z_{ji} is the j^{th} confounder of the i^{th} exposure.
- ε_i is the residual of the expectation of the i^{th} response conditional on the i^{th} exposure and its related confounders.

The standard error and 95% confidence interval will be obtained by the bootstrap method based on 10,000 samplings with replacement. As for the p-value, it will be tested according to the following hypothesis:

- H_0: The causal effect of the exposure on the outcome is equal to 0.
- H_1: The causal effect of the exposure on the outcome is not equal to 0.

We need to introduce another boostrap, which is constructed under H_0, to assist the test-statistic computation. This bootstrap will also be made up of 10,000 samplings with replacement. Before running the bootstrap, we will permute the exposures randomly amongst participants who attend 18-years measurement wave and have value in the interested exposure. The purpose of this step is to break the causal relationship between the exposure and the final outcome to construct the distribution under H_0 for the test-statistic. This method will give a robust p-value as it accounts for the variability incurred by estimating the regression weights (Details in 2.5). The test-statistic will be obtained by comparing the t-statistics from the bootstrap samples under the null distribution and the observed t-statistics from the original data. The formula is:

$$p - value = \frac{\#[n_{t_{boot}<-|t_{obs}|} + n_{t_{boot}>|t_{obs}|}]}{\#[n_s]}$$

where

- $\#[n_{t_{boot}<-|t_{obs}|}]$ is the number of t-statistic from the bootstrap samples under the null distribution less than the negative absolute value of the observed t-statistic from the original data.
- $\#[n_{t_{boot}>|t_{obs}|}]$ is the number of t-statistic from the bootstrap samples under the null distribution greater than the absolute value of the observed t-statistic from the original data.
- $\#[n_s]$ is the number of bootstrap samples.

Based on the formula, the p-value will be significant if most of t-statistics in the bootstrap samples under the null distribution do not exceed, in magnitude, the observed t-statistics from the original data as obtained under the alternative. The observed t-statistic will then be in the tail end of the null distribution in this scenario. This means that it is not compatible with H_0. Therefore, we will have enough evidence to reject H_0 and statistically accept H_1 if the p-value is significant. Typically, the larger the absolute value of t-statistics is, the smaller the p-value is and the less consistent is the data from H_0. This is because a larger t-statistics indicates that the estimated causal effect is considerably different from 0 compared to its standard error.

2.5 Selection bias

There is often selection bias in cohort study data. This is due to it being an observation study. A cohort study cannot achieve proper randomization to eliminate the bias introduced by selecting the interested observations. This usually refers to the bias introduced by conditioning on the common effects, such as informative drop-out, non-response/missing data, and self-selection [42]. These common effects are not able to be completely avoided in the cohort study. In our case, the common effects we need to account for are: 1. informative drop-out: some of the children did not attend the 18-years measurement wave, for reasons that may involve exposure or outcome; 2. non-response/missing data: some of the children did not answer any questions in the food questionnaires although they attended the measurement wave. Selection bias is be taken into account when building the model. The analysis will draw an incorrect conclusion If we do not take any action to fix selection bias.

We implemented Inverse Probability Weight (IPW) [42] to reduce the impact of selection bias on the causal model. This method can re-balance the dataset by up-weighting underrepresented observations. In the analysis, absent participants at 18 years observations are thus represented by the participants with similar characters. To do so, we estimated the probability for each child in the cohort supporting complete data by regressing an indicator of completeness on selected variables from the birth cohort, using logistic regression. We then used its inverse as the weighting for all participants seen at 18 years. We set up 16 complete data indicators. They correspond to each food category (4 at each measurement wave) at 4 measurement waves respectively. Each complete data indicators represents whether a child answered questions related to a specified eating pattern (e.g. occasional eating pattern) in the food questionnaire at the particular measurement wave (e.g. at 4-years wave) and whether they attended the 18-years measurement wave. The details of selected variables from the birth cohort is shown in Table 2.7.

Table 2.7: List of selected variables from birth cohort for Inverse Probability Weighting (IPW)

Domain	Age	Variable Name	Type	Unit	Note
General	6 weeks	t0p_samoan_rec	categorical	-	An indicator shows whether neither mother nor father is Samoan, or either mother or father is Samoan, or both mother and father is Samoan
General	6 weeks	t0p_cook_rec	categorical	-	An indicator shows whether neither mother nor father is Cook Island, or either mother or father is Cook Island, or both mother and father is Cook Island
General	6 weeks	t0p_tongan_rec	categorical	-	An indicator shows whether neither mother nor father is Tongan, or either mother or father is Tongan, or both mother and father is Tongan

Continued on next page

Table 2.7 – *Continued from previous page*

Domain	Age	Variable Name	Type	Unit	Note
General	6 weeks	t0p_otherp_rec	categorical	-	An indicator shows whether neither mother nor father is Other Pacific Island, or either mother or father is Other Pacific Island, or both mother and father is Other Pacific Island
General	6 weeks	t0p_othernp_rec	categorical	-	An indicator shows whether neither mother nor father is Other Non Pacific Island, or either mother or father is Other Non Pacific Island, or both mother and father is Other Non Pacific Island
General	6 weeks	t0pa4	categorical	-	Baby's gender
General	6 weeks	t0pi1	binary	-	With both natural parents
General	6 weeks	t0pi2	binary	-	With adoptive parents
General	6 weeks	t0pi3	binary	-	With a single parent family
General	6 weeks	t0pi4	binary	-	With a step parent
General	6 weeks	t0pi5	binary	-	In another relative's home
General	6 weeks	t0pi6	binary	-	In a foster family
General	6 weeks	t0pi9	binary	-	Other (specify)

Continued on next page

Table 2.7 – *Continued from previous page*

Domain	**Age**	**Variable Name**	**Type**	**Unit**	**Note**
General	6 weeks	t0pl1	categorical	-	Highest school qualification
General	6 weeks	t0pl2	categorical	-	Highest post-school qualification
General	6 weeks	t0pl3	categorical	-	Employment situation prior to pregnancy
General	6 weeks	l3_rec	binary	-	Mother employed prior to pregnancy (T0PL3 recode)
General	6 weeks	l7_rec	binary	-	Mothers employment (T0PL7 recode)
General	6 weeks	inc_cat2	categorical	-	Household income: 4 categories
General	6 weeks	agecat2	categorical	-	Age: 4 groups
Immunisation	6 weeks	immun	binary	-	immunised at 6 weeks
Breastfeeding	6 weeks	d1_recod	categorical	-	How fed baby 1st 6 wks (D1 recoded 3=1)
Antenatal smoking	6 weeks	j23_25re	binary	-	Smoked during pregnancy
Antenatal smoking	6 weeks	t0p_pgcigs_rec	continuous	stick	# cigarettes during pregnancy; Add, multiply by 91/365, divide by 20 to obtain the pack-years of smoking during pregnancy

Continued on next page

Table 2.7 – *Continued from previous page*

Domain	**Age**	**Variable Name**	**Type**	**Unit**	**Note**
Smoking exposure	6 weeks	t0p_liv_rec	binary	-	# living in household now who smoke; Dichotomise 0 as ”N” and non-zero as ”Y”
Respiratory illness-infection	6 weeks	t0pc20	binary	-	Problems with breathing
Weight, height, BMI	6 weeks	t0h12	continuous	gram	Birth weight (gms)
Dwelling	6 weeks	t0p_damp_rec	binary	-	Dampness/mould. Dichotomise
Dwelling	6 weeks	t0p_cold_rec	binary	-	Cold. Dichotomise
Dwelling	6 weeks	t0p_oc_rec	binary	-	Overcrowding. Dichotomise

2.6 Population attributable fraction

Population attributable fraction (PAF) is a popular epidemiology measure of the proportion of a specified outcome that is attributable to the effects of a particular modifiable risk or a protective factor at the population level [43]. To promote strength-based interpretation, we will be considering the PAF of healthy lung function associated with each nutrition factor. If the nutrition exposure were binary in nature, then

$$PAF = N(1,0) = \frac{N_1 - N_0}{N_1}$$

where

- N_1 is the current prevalence of healthy lung function.
- N_0 is the same prevalence at a nutrition exposure of 0 portions per day.

To accommodate a continuous exposure, we follow the methods outlined in [44], with a natural floor (or minimum risk exposure value) of 0 portions per day. In this instance, for any value x of the nutrition factor score, we can adapt Equation 7 from [43] to obtain

$$PAF = \int_{x,z} p(x,z|y=1)[1 - \frac{1}{rr(x,z)}]dzdx$$

where

- z is the set of confounders.
- y is the indicator of healthy lung function.
- $rr(x,z)$ is the relative risk of healthy lung function under confounder set z between x and 0 portions per day.

Under the relative risk regrssion model $P[y=1|x,z] = e^{(\alpha+\beta x+\sum_{j=1}^{k}\ \gamma_j z_j)}$, it can be seen that $rr(x,z) = e^{\beta x}$, and that the PAF can be consistently estimated by

$$P\hat{A}F = \frac{1}{W}\sum_{i=1,y=1}^{n} w_i[1 - e^{-\hat{\beta}x_i}]$$

where

- w_i is inverse Probability Weighting (IPW) for the i_th observation.
- $W = \sum_{i=1,y=1}^{n} w_i$.

- $\hat{\beta}$ is the estimated causal effect.
- x_i is the factor score of the i_th observation.

It is able to tell us the proportion of the current risk event that will be avoided by eliminating a risk factor of interest or will increase by removing a protective factor of interest. Therefore, we can gain the priorities for intervening in the risk factors or protective factors from the PAF so as to minimize the occurrence of the specified risk event. In this paper, we hope to gain the PAF of nutrition factor scores on the lung function which is conditional on the confounders. This PAF will be computed based on the result of the relative risk model. Its equation is [44]:

$$PAF = (\sum_{i=1}^{n} w_i)^{-1} \sum_{i=1}^{n} [w_i(1 - e^{-\beta x_i})]$$

where

- w_i is inverse Probability Weighting (IPW) for the i_th observation.
- β is the estimated causal effect.
- x_i is the factor score of the i_th observation.

This formula also allows us to investigate the effect of changes to the distribution of the exposure on healthy lung function prevalence. As a simple example of such an investigation, we added constant offsets to all x_i under while remaining consistent with the original range of the related eating pattern. In order to achieve this, we used the lower bound of the original range instead of it if the shifted value was less than the lower bound, and used the upper bound of the original range instead of the shifted value if it was larger than the upper bound. We carried out this exercise over a range of offsets to illustrate how the effect of a public health intervention on nutrition might affect the PAF of nutrition on healthy lung function. We also calculated an indicative 95% confidence interval for the modified location model by simply shifting by the offset value the 95% confidence bounds at the zero-offset location, obtained from the bootstrap method.

Chapter 3

Data analysis

3.1 Selected variables

Based on the discussion with the research team, we selected the most relevant variables from each domain. The list of selected variables is shown in Table 3.1, and descriptive table is given in Table 3.2 and Table 3.3.

Table 3.1: List of selected variables

Domain	Age	Variable Name	Type	Unit	Note
Immunisation	6 weeks - 1 year	t1p_imm_rec (Immun 5m)	categorical	-	Combine immunised at 6 weeks, 3 month immunisation, 5 months immunisation to ”None, Partial, or Complete”
Immunisation	2 years	t2pf230 (Immun 15m)	binary	-	15 month immunisations
Immunisation	4 years	t4pf66 (Immun 4-5y)	binary	-	4-5 yrs immunisations
Immunisation	6 years	t6p_men_rec (Immun mening 6y)	binary	-	Transform immunisation1 and immunisation2 to whether a child gets Meningococcal
Immunisation	6 years	t6pf53 (Immun std 6y)	binary	-	Standard age-related immunisation
Exercise	4 years	t4ch_pa_rec (Activity time 4y)	continuous	hour	days*(hours + mins/60) per week in moderate and vigorous activity
Exercise	11 years	t11a_pa_rec (Activity 11y)	binary	-	Combine Never, Less than once and Once a week to ”N”, and combine Several and Every day to ”Y”

Continued on next page

Table 3.1 – *Continued from previous page*

Domain	Age	Variable Name	Type	Unit	Note
Exercise	14 years	t14p_mumpa_rec (Activity 14y)	binary	-	Combine 2 or fewer to ”N”, and combine 3 or more to ”Y”
Exercise	14 years	t14p_ipa_rec (Activity incid 14y)	binary	-	Provisionally, add the number of days of ”To” and ”From” (Incidental physical activity) then dichotomise 0 as ”N” and non-zero as ”Y”
Exercise	14 years	t14p_ppa_rec (Activity purpose 14y)	binary	-	Add the number of days with purposeful physical activity in weekdays and weekends then dichotomise 0-2 days as ”N” and 3 or more as ”Y”
Breastfeeding	6 weeks	d1_recod (Feeding 6wk)	categorical	-	How fed baby 1st 6 wks (D1 recoded 3=1)
Breastfeeding	2 years	t2pc16 (Feeding 2y)	binary	-	Feeding at 2 years
Antenatal smoking	6 weeks	j23_25re (Smoked preg)	binary	-	Smoked during pregnancy

Continued on next page

Table 3.1 – *Continued from previous page*

Domain	Age	Variable Name	Type	Unit	Note
Antenatal smoking	6 weeks	t0p_pgcigs_rec (Smoke preg pack-y)	continuous	stick	# cigarettes during pregnancy; Add, multiply by 91/365, divide by 20 to obtain the pack-years of smoking during pregnancy
Smoking exposure	6 weeks	t0p_liv_rec (Smoker in dwelling 6wk)	binary	-	# living in household now who smoke. Dichotomise 0 as ”N” and non-zero as ”Y”
Smoking exposure	1 year	t1p_liv_rec (Smoker in dwelling 1y)	binary	-	# living in household now who smoke. Dichotomise 0 as ”N” and non-zero as ”Y”
Smoking exposure	2 years	t2p_liv_rec (Smoker in dwelling 2y)	binary	-	# living in household now who smoke. Dichotomise 0 as ”N” and non-zero as ”Y”
Smoking exposure	4 years	t4p_liv_rec (Smoker in dwelling 4y)	binary	-	# living in household now who smoke. Dichotomise 0 as ”N” and non-zero as ”Y”

Continued on next page

Table 3.1 – *Continued from previous page*

Domain	**Age**	**Variable Name**	**Type**	**Unit**	**Note**
Smoking exposure	6 years	t6p_liv_rec (Smoker in dwelling 6y)	binary	-	# living in household now who smoke. Dichotomise 0 as ”N” and non-zero as ”Y”
Smoking exposure	9 years	t9p_liv_rec (Smoker in dwelling 9y)	binary	-	# living in household now who smoke. Dichotomise 0 as ”N” and non-zero as ”Y”
Smoking exposure	11 years	t11p_liv_rec (Smoker in dwelling 11y)	binary	-	# living in household now who smoke. Dichotomise 0 as ”N” and non-zero as ”Y”
Smoking exposure	14 years	t14p_liv_rec (Smoker in dwelling 14y)	binary	-	# living in household now who smoke. Dichotomise 0 as ”N” and non-zero as ”Y”
Smoking	14 years	t14a_smokingstatus (Smoke status 14y)	binary	-	-
Respiratory illness-infection	6 weeks	t0pc20 (Breath prob 6wk)	binary	-	Problems with breathing
Respiratory illness-infection	1 year	t1p_dib_rec (Breath prob 1y)	binary	-	Problems with breathing. Dichotomise 0 against more than 0.

Continued on next page

Table 3.1 – *Continued from previous page*

Domain	Age	Variable Name	Type	Unit	Note
Respiratory illness-infection	2 years	t2p_dib_rec (Breath prob 2y)	binary	-	Problems with breathing. Dichotomise 0 against more than 0.
Respiratory illness-infection	4 years	t4p_dib_rec (Breath prob 4y)	binary	-	Dichotomise t4pf3 is equal to ”Yes”, t4pf4 is equal to ”Sometimes” or ”Often”, and t4pf5 is equal to ”Sometimes” or ”Often” as ”Y” against others as ”N”
Respiratory illness-infection	4 years	t4pf6 (Asthma Dx 4y)	binary	-	Child ever diagnosed with asthma
Respiratory illness-infection	6 years	t6p_dib_rec (Breath prob 6y)	binary	-	Problems with breathing. Dichotomise 0 against more than 0.
Respiratory illness-infection	6 years	t6pf7 (Asthma Dx 6y)	binary	-	Child ever diagnosed with asthma
Respiratory illness-infection	9 years	t9pf32 (Asthma Dx 9y)	binary	-	Child ever diagnosed with asthma
Respiratory illness-infection	11 years	t11pe11 (Asthma Dx 11y)	binary	-	Child ever diagnosed with asthma
Weight, height, BMI	6 weeks	t0h12 (Weight birth)	continuous	gram	Birth weight (gms)
Weight, height, BMI	1 year	t1pb26 (Weight 1y)	continuous	gram	Weight (gms)

Continued on next page

Table 3.1 – *Continued from previous page*

Domain	**Age**	**Variable Name**	**Type**	**Unit**	**Note**
Weight, height, BMI	2 years	t2ch_bmi_rec (BMI 2y)	continuous	-	Combine Weight (kg) and Height (cm) to form BMI
Weight, height, BMI	4 years	t4ch_bmi_rec (BMI 4y)	continuous	-	Combine Weight (kg) and Height (cm) to form BMI
Weight, height, BMI	6 years	t6ch_bmi (BMI 6y)	continuous	-	-
Weight, height, BMI	9 years	t9ch_bmi_rec (BMI 9y)	continuous	-	Combine Weight (kg) and Height (cm) to form BMI
Weight, height, BMI	11 years	t11ch_bmi_rec (BMI 11y)	continuous	-	Combine Weight (kg) and Height (cm) to form BMI
Weight, height, BMI	14 years	t14ch_bmi_rec (BMI 14y)	continuous	-	Combine Weight (kg) and Height (cm) to form BMI
Allergies	9 years	t9pf1 (Allergies 9y)	binary	-	Does child have allergies
Dwelling	6 weeks	t0p_damp_rec (Dwelling damp 6wk)	binary	-	Dampness/mould. Dichotomise
Dwelling	6 weeks	t0p_cold_rec (Dwelling cold 6wk)	binary	-	Cold. Dichotomise

Continued on next page

Table 3.1 – *Continued from previous page*

Domain	**Age**	**Variable Name**	**Type**	**Unit**	**Note**
Dwelling	6 weeks	t0p_oc_rec (Dwelling overcrowding 6wk)	binary	-	Overcrowding. Dichotomise
Dwelling	1 year	t1p_damp_rec (Dwelling damp 1y)	binary	-	Dampness/mould. Dichotomise
Dwelling	1 year	t1p_cold_rec (Dwelling cold 1y)	binary	-	Cold. Dichotomise
Dwelling	1 year	t1p_oc_rec (Dwelling overcrowding 1y)	binary	-	Overcrowding. Dichotomise
Dwelling	2 years	t2p_damp_rec (Dwelling damp 2y)	binary	-	Combine with dampness and mould.
Dwelling	2 years	t2p_cold_rec (Dwelling cold 2y)	binary	-	Cold. Dichotomise
Dwelling	2 year	t2p_oc_rec (Dwelling overcrowding 2y)	binary	-	Overcrowding. Dichotomise
Dwelling	14 years	t14cq7d (Dwelling damp 14y)	binary	-	It’s damp
Dwelling	14 years	t14cq7e (Dwelling cold 14y)	binary	-	It’s too cold or difficult to heat/ keep warm

Table 3.2: Descriptive table for selected variables (continuous)

Variable	**n**	**miss**	**p.miss**	**mean**	**sd**	**median**	**p25**	**p75**	**min**	**max**
Exercise										
t4ch_pa_rec (Activity time 4y)	1,398	507	36.3	30	20	20	10	40	0	200
Antenatal smoking										
t0p_pgcigs_rec (Smoke preg pack-y)	1,398	6	0.4	0.07	0.2	0	0	0	0	2
Weight, height, BMI										
t0h12 (Weight birth)	1,398	17	1.2	4,000	600	4,000	3,000	4,000	600	5,000
t1pb26 (Weight 1y)	1,398	170	12.2	10,000	2,000	10,000	10,000	10,000	5,000	20,000
t2ch_bmi_rec (BMI 2y)	1,398	369	26.4	20	2	20	20	20	10	40
t4ch_bmi_rec (BMI 4y)	1,398	514	36.8	20	2	20	20	20	10	40
t6ch_bmi (BMI 6y)	1,398	505	36.1	20	3	20	20	20	10	40
t9ch_bmi_rec (BMI 9y)	1,398	514	36.8	20	5	20	20	30	10	40
t11ch_bmi_rec (BMI 11y)	1,398	450	32.2	20	5	20	20	30	10	50
t14ch_bmi_rec (BMI 14y)	1,398	485	34.7	30	7	30	20	30	20	60

Note: n - the total number of observations; miss - the number of observations with missing value; p.miss - the percentage of observations with missing value accounted for the total number; sd - standard deviation; p25 - 25% percentile; p75 - 75% percentile

Table 3.3: Descriptive table for selected variables (binary & categorical)

Variable	**n**	**level**	**freq**	**percent**	**cum.percent**
Immunisation					
t1p_imm_rec	1,398	Complete	876	62.7	62.7
(Immun 5m)		None	46	3.3	66.0
		Partial	476	34.0	100.0
t2pf230	1,398	No	94	6.7	6.7
(Immun 15m)		Yes	1,065	76.2	82.9
		<NA>	239	17.1	100.0
t4pf66	1,398	No	626	44.8	44.8
(Immun 4-5y)		Yes	439	31.4	76.2
		<NA>	333	23.8	100.0
t6p_men_rec	1,398	No	14	1.0	1.03
(Immun mening 6y)		Yes	722	51.6	52.6
		<NA>	662	47.4	100.0
t6pf53	1,398	No	23	1.6	1.6
(Immun std 6y)		Yes	994	71.1	72.7
		<NA>	381	27.3	100.0
Exercise					
t11a_pa_rec	1,398	No	299	21.4	21.4
(Activity 11y)		Yes	650	46.5	67.9
		<NA>	449	32.1	100.0
t14p_mumpa_rec	1,398	No	464	33.2	33.2
(Activity 14y)		Yes	478	34.2	67.4
		<NA>	456	32.6	100.0
t14p_ipa_rec	1,398	No	344	24.6	24.6
(Activity incid 14y)		Yes	489	35.0	59.6
		<NA>	565	40.4	100.0
t14p_ppa_rec	1,398	No	171	12.2	12.2
(Activity purpose 14y)		Yes	633	45.3	57.5
		<NA>	594	42.5	100.0

Continued on next page

Table 3.3 – *Continued from previous page*

Variable	n	level	freq	percent	cum.percent
Breastfeeding					
d1_recod	1,398	Only with breast milk	686	49.1	49.1
(Feeding 6wk)		Combination breast milk and other	532	38.1	87.1
		Only with formula or other bottle milk	180	12.9	100.0
t2pc16	1,398	No	990	70.8	70.8
(Feeding 2y)		Yes	171	12.2	83.0
		<NA>	237	17.0	100.0
Antenatal smoking					
j23_25re	1,398	No	1,047	74.9	74.9
(Smoked preg)		Yes	345	24.7	99.6
		<NA>	6	0.4	100.0
Smoking exposure					
t0p_liv_rec	1,398	No	682	48.8	48.8
(Smoker in dwelling 6wk)		Yes	713	51.0	99.8
		<NA>	3	0.2	100.0
t1p_liv_rec	1,398	No	651	46.6	46.6
(Smoker in dwelling 1y)		Yes	585	41.8	88.4
		<NA>	162	11.6	100.0
t2p_liv_rec	1,398	No	627	44.8	44.8
(Smoker in dwelling 2y)		Yes	527	37.7	82.5
		<NA>	244	17.5	100.0
t4p_liv_rec	1,398	No	567	40.6	40.6
(Smoker in dwelling 4y)		Yes	493	35.3	75.8
		<NA>	338	24.2	100.0
t6p_liv_rec	1,398	No	601	43.0	43.0
(Smoker in dwelling 6y)		Yes	409	29.3	72.2
		<NA>	388	27.8	100.0
t9p_liv_rec	1,398	No	587	42.0	42.0
(Smoker in dwelling 9y)		Yes	382	27.3	69.3
		<NA>	429	30.7	100.0
t11p_liv_rec	1,398	No	604	43.2	43.2
(Smoker in dwelling 11y)		Yes	441	31.5	74.7
		<NA>	353	25.3	100.0

Continued on next page

Table 3.3 – *Continued from previous page*

Variable	n	level	freq	percent	cum.percent
t14p_liv_rec	1,398	No	561	40.1	40.1
(Smoker in dwelling 14y)		Yes	389	27.8	68.0
		<NA>	448	32.0	100.0
Smoking					
t14a_smokingstatus	1,398	No	845	60.4	60.4
(Smoke status 14y)		Yes	60	4.3	64.7
		<NA>	493	35.3	100.0
Respiratory illness-infection					
t0pc20	1,398	No	799	57.2	57.2
(Breath prob 6wk)		Yes	598	42.8	99.9
		<NA>	1	0.1	100.0
t1p_dib_rec	1,398	No	176	12.6	12.6
(Breath prob 1y)		Yes	1,065	76.2	88.8
		<NA>	157	11.2	100.0
t4p_dib_rec	1,398	No	797	57.0	57.0
(Breath prob 4y)		Yes	245	17.5	74.5
		<NA>	356	25.5	100.0
t4pf6	1,398	No	869	62.2	62.2
(Asthma Dx 4y)		Yes	185	13.2	75.4
		<NA>	344	24.6	100.0
t6p_dib_rec	1,398	No	718	51.4	51.4
(Breath prob 6y)		Yes	300	21.5	72.8
		<NA>	380	27.2	100.0
t6pf7	1,398	No	886	63.4	63.4
(Asthma Dx 6y)		Yes	133	9.5	72.9
		<NA>	379	27.1	100.0
t9pf32	1,398	No	845	60.4	60.4
(Asthma Dx 9y)		Yes	168	12.0	72.5
		<NA>	385	27.5	100.0
t11pe11	1,398	No	878	62.8	62.8
(Asthma Dx 11y)		Yes	169	12.1	74.9
		<NA>	351	25.1	100.0

Continued on next page

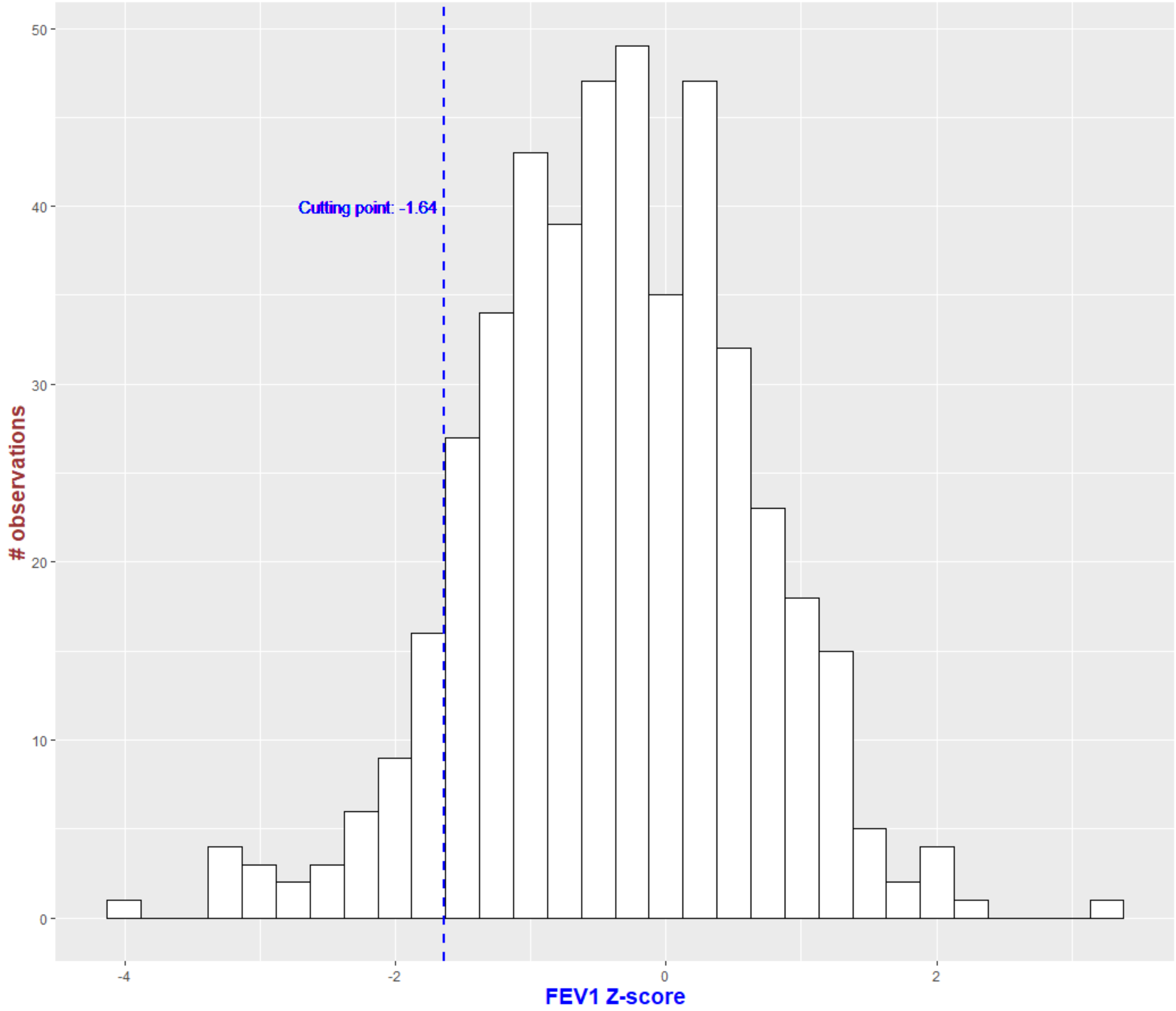
Table 3.3 – *Continued from previous page*

Variable	n	level	freq	percent	cum.percent
Allergies					
t9pf1	1,398	No	863	61.7	61.7
(Allergies 9y)		Yes	150	10.7	72.5
		<NA>	385	27.5	100.0
Dwelling					
t0p_damp_rec	1,398	No	873	62.4	62.4
(Dwell damp 6wk)		Yes	522	37.3	99.8
		<NA>	3	0.2	100.0
t0p_cold_rec	1,398	No	645	46.1	46.1
(Dwell cold 6wk)		Yes	752	53.8	99.9
		<NA>	1	0.1	100.0
t1p_damp_rec	1,398	No	559	40.0	40.0
(Dwell damp 1y)		Yes	682	48.8	88.8
		<NA>	157	11.2	100.0
t1p_cold_rec	1,398	No	447	32.0	32.0
(Dwell cold 1y)		Yes	794	56.8	88.8
		<NA>	157	11.2	100.0
t2p_damp_rec	1,398	No	362	25.9	25.9
(Dwell damp 2y)		Yes	799	57.2	83.0
		<NA>	237	17.0	100.0
t2p_cold_rec	1,398	No	483	34.5	34.5
(Dwell cold 2y)		Yes	679	48.6	83.1
		<NA>	236	16.9	100.0
t14cq7d	1,398	No	571	40.8	40.8
(Dwell damp 14y)		Yes	75	5.4	46.2
		<NA>	752	53.8	100.0
t14cq7e	1,398	No	494	35.3	35.3
(Dwell cold 14y)		Yes	152	10.94	46.2
		<NA>	752	53.8	100.0

Note: n - the total number of observations; freq - the number of observations in the level; percent - the percentage of observations in the level accounted for the total number; cum.percent - the cumulative percentage; <NA> - No response

Figure 3.1 presents the distribution of FEV1 Z-score, and the cutting point (-1.64) which is used in the relative risk model. The cutting point is located around 10^{th}centile of the empirical distribution.

Figure 3.1: The distribution of FEV1 Z-score, and cutting point used in the relative risk model (-1.64)



3.2 Dimensional reduction over Nutrition data

3.2.1 Results from Exploratory Factor Analysis (EFA)

Table 3.4 presents the eating patterns and related factor loadings published in [28]. Four eating patterns were therein established based on the data from the 14-years measurement wave. They are respectively identified by using different colours in this paper: Blue - ”Occasional” (Eating pattern 1), Green - ”Seafood” (Eating pattern 2), Yellow - ”Fruit and vegetables” (Eating pattern 3 - renamed from ”Basic and staples”), and Red - ”Meat and bread” (Eating pattern 4). We also obtained the factorial structure of the food categories in the selected measurement waves (4 years, 6 years, 9 years, and 14 years) by Exploratory Factor Analysis (EFA). Details are shown in Table 3.5, Table 3.6, Table 3.7, and Table 3.8. In these tables, the food categories were colored similarly to the original eating patterns even though they were allocated to another eating pattern.

There are some noticeable points when comparing the eating patterns generated by EFA at different ages. The eating patterns are quite similar between 9-years and 14-years measurement wave. The only difference is that eating pattern 4 at 9 years does not load on ”Bread/ toast/ bread rolls”. However, the eating patterns at 4 years are somehow different from the eating patterns at 9 years and 14 years. The differences are revealed in eating patterns 1, 2, and 4. In eating pattern 1, ”Fruit juices and fruit drinks” is removed and ”Red meat” and ”Processed meat products” are added. At the same time, eating pattern 2 and 4 only have loadings from ”Fruit juices and fruit drinks” and ”Chicken” respectively. By contrast, the eating patterns at 6 years have few similarities to the eating patterns at other ages except the eating pattern 1. Eating pattern 3 is the most unusual eating pattern at 6 years. It loads on ”Fruit (fresh/ frozen/ canned/ stewed)” and ”Chicken”. Lastly, It is worth noting that none of the eating patterns at 4 years and 6 years load onto from ”Canned fish/ shellfish” and ”Fresh/ frozen fish/ shellfish”.

We can compare the eating results obtained from EFA to the original results according to the colored labels. First, the eating patterns at 9 years and 14 years obtained from EFA are almost identical to the original results. Furthermore, 2 of the 4 eating patterns at 4 years are similar to these results, and some food categories are still occurring together: 1.”Fast food/ takeaways”, ”Soft drinks/ energy drinks”, and ”Lollies, sweets, chocolate and confectionery” (Blue); 2. ”Red meat” and ”Processed meat products” (Red); 3. ”Vegetables (fresh/ frozen/ canned)” and ”Fruit (fresh/ frozen/ canned/ stewed)” (Yellow). However, the combination 2 is loaded on the eating pattern 1 rather than the eating pattern 4 at 4 years. In contrast, the eating patterns at 6 years are considerably different from the original eating patterns, and not many food categories are still occurring together. The eating patterns in 3 of the 4 measurement waves are consistent with the eating patterns originally defined. Therefore, this thesis aligns itself with the original eating patterns. Another thing to note is that the factor loadings at 14 years in this thesis are different from the factor loadings in the original thesis. This has occurred is principally because the defined food categories are not identical the same between these two papers, having needed to remain consistent across measurement waves for our purposes.

Table 3.4: Original eating patterns based on 14-years measurement wave [28]

Food category	**Eating Pattern 1**	**Eating Pattern 2**	**Eating Pattern 3**	**Eating Pattern 4**
Fast food/ takeaways	🟦			
Soft drinks/ energy drinks	🟦			
Lollies, sweets, chocolate and confectionery	🟦			
Fruit juices and fruit drinks	🟦			
Canned fish/ shellfish		🟩		
Fresh/ frozen fish/ shellfish		🟩		
Vegetables (fresh/ frozen/ canned)			🟨	
Fruit (fresh/ frozen/ canned/ stewed)			🟨	
Red meat				🟥
Chicken				🟥
Processed meat products				🟥
Bread/ toast/ bread rolls				🟥
Name of eating pattern	Occasional	Seafood	Fruit and vegetables	Meat and bread

Note: Hot chips, French fries, wedges/ kumara chips, Battered/ fired fish/ shellfish, and Milk were excluded

Table 3.5: EFA factor loadings at 14-years

Food category	**Eating Pattern 1**	**Eating Pattern 2**	**Eating Pattern 3**	**Eating Pattern 4**
Fast food/ takeaways	0.776			
Soft drinks/ energy drinks	1.210			
Lollies, sweets, chocolate and confectionery	1.038			
Fruit juices and fruit drinks	0.492			
Canned fish/ shellfish		1.270		
Fresh/ frozen fish/ shellfish		0.426		
Vegetables (fresh/ frozen/ canned)			1.235	
Fruit (fresh/ frozen/ canned/ stewed)			1.046	
Red meat				0.904
Chicken				1.222
Processed meat products				1.066
Bread/ toast/ bread rolls				0.476

Table 3.6: EFA factor loadings at 9-years

Food category	**Eating Pattern 1**	**Eating Pattern 2**	**Eating Pattern 3**	**Eating Pattern 4**
Fast food/ takeaways	0.435			
Soft drinks/ energy drinks	0.742			
Lollies, sweets, chocolate and confectionery	0.569			
Fruit juices and fruit drinks	0.474			
Canned fish/ shellfish		0.540		
Fresh/ frozen fish/ shellfish		0.514		
Vegetables (fresh/ frozen/ canned)			0.954	
Fruit (fresh/ frozen/ canned/ stewed)			0.428	
Red meat				0.430
Chicken				0.418
Processed meat products				0.400
Bread/ toast/ bread rolls				

Table 3.7: EFA factor loadings at 6-years

Food category	**Eating Pattern 1**	**Eating Pattern 2**	**Eating Pattern 3**	**Eating Pattern 4**
Fast food/ takeaways	0.824			
Soft drinks/ energy drinks		0.668		
Lollies, sweets, chocolate and confectionery	0.532			
Fruit juices and fruit drinks		0.484		
Canned fish/ shellfish				
Fresh/ frozen fish/ shellfish				
Vegetables (fresh/ frozen/ canned)				0.409
Fruit (fresh/ frozen/ canned/ stewed)			0.698	
Red meat	0.654			
Chicken			0.585	
Processed meat products	0.485			
Bread/ toast/ bread rolls				

Table 3.8: EFA Factor loadings at 4-years

Food category	**Eating Pattern 1**	**Eating Pattern 2**	**Eating Pattern 3**	**Eating Pattern 4**
Fast food/ takeaways	0.712			
Soft drinks/ energy drinks	0.564			
Lollies, sweets, chocolate and confectionery	0.492			
Fruit juices and fruit drinks		0.997		
Canned fish/ shellfish				
Fresh/ frozen fish/ shellfish				
Vegetables (fresh/ frozen/ canned)			0.913	
Fruit (fresh/ frozen/ canned/ stewed)			0.401	
Red meat	0.612			
Chicken				0.438
Processed meat products	0.635			
Bread/ toast/ bread rolls				

3.2.2 Results from Confirmatory Factor Analysis (CFA)

Following the discussion from the previous subsection, we fitted Confirmatory Factor Analysis (CFA) models based on the eating patterns published in [28], and tested their measurement invariance. In the first run, we found that ”Bread/ toast/ bread rolls” uniformly did not perform well in both of the CFA models (R-squared: From 0.006 to 0.073 in configural invariance model and from 0.003 to 0.019 in structural invariance model). This occurs since there is more missing data in ”Bread” category. Therefore, we determined to remove this food category from the models and renamed the ”Meat and bread” eating pattern to the ”Meat” eating pattern. ”Chicken” also did not perform well in the structural invariance model but there were some reasonable R-squared shown in the configural invariance model. We decided to keep it in CFA model.

According to R-squared criterion, the structural invariance model accounts for more variances at 6 years and 9 years. In terms of food categories, this model has a better performance in ”Fast food/ takeaways”, ”Processed meat products”, and ”Vegetables (fresh/ frozen/ canned)” but have a poor performance in ”Chicken”. The full list of R-squared is shown in Table 3.9.

Table 3.9: R square for food categories in the structural invariance model

Food category	**R square**			
	4 years	6 years	9 years	14 years
Fast food/ takeaways	0.424	0.608	0.477	0.372
Soft drinks/ energy drinks	0.094	0.235	0.513	0.305
Lollies, sweets, chocolate and confectionery	0.170	0.263	0.513	0.423
Fruit juices and fruit drinks	0.051	0.166	0.255	0.251
Canned fish/ shellfish	0.081	0.300	0.182	0.137
Fresh/ frozen fish/ shellfish	0.110	0.417	0.387	0.208
Vegetables (fresh/ frozen/ canned)	0.472	0.410	0.386	0.369
Fruit (fresh/ frozen/ canned/ stewed)	0.261	0.496	0.340	0.249
Red meat	0.187	0.395	0.531	0.427
Chicken	0.017	0.070	0.029	0.027
Processed meat products	0.186	0.546	0.576	0.576

Table 3.10 shows factor loadings generated by the structural invariance model. As we can see, there is no strong dominant food category in the ”Occasional” eating pattern. This pattern loads at 36% on ”Fast food/ takeaways”, 24% on ”Lollies, sweets, chocolate and confectionery”, 22% on ”Soft drinks/ energy drinks”, and 19% on ”Fruit juices and fruit drinks”. Furthermore, the ”Seafood” eating pattern is weakly dominated by ”Fresh/ frozen fish/ shellfish” and the ”Fruit and vegetables” eating pattern by ”Vegetables (fresh/ frozen/ canned)”. The ”Meat” eating pattern is strongly dictated by ”Red meat” and ”Processed meat product”. The factor score computation approach we selected does not maintain the correlation structure of the factors amongst the factor scores, as shown in Tables 3.11 to 3.15. Meanwhile, none of correlation matrix of factor scores indicates any considerably strong correlations.

Table 3.10: CFA factor loadings based on the structural invariance model

Food category	**Occasional**	**Seafood**	**Fruit and vegetables**	**Meat**	**Normalised loading**
Fast food/ takeaways	1.000				0.355
Soft drinks/ energy drinks	0.611				0.217
Lollies, sweets, chocolate and confectionery	0.674				0.240
Fruit juices and fruit drinks	0.529				0.188
Canned fish/ shellfish		1.000			0.421
Fresh/ frozen fish/ shellfish		1.374			0.579
Vegetables (fresh/ frozen/ canned)			1.000		0.543
Fruit (fresh/ frozen/ canned/ stewed)			0.852		0.457
Red meat				1.000	0.468
Chicken				0.172	0.078
Processed meat products				0.971	0.454

Table 3.11: Correlation matrix of factors (identical across all measurement waves)

	Occasional	**Seafood**	**Fruit and vegetables**	**Meat**
Occasional	1.000			
Seafood	0.569	1.000		
Fruit and vegetables	0.187	0.522	1.000	
Meat	0.641	0.730	0.614	1.000

Table 3.12: Correlation matrix of factor scores based on eating patterns at 4 years

	Occasional	**Seafood**	**Fruit and vegetables**	**Meat**
Occasional	1.000			
Seafood	0.303	1.000		
Fruit and vegetables	0.341	0.310	1.000	
Meat	0.450	0.374	0.382	1.000

Table 3.13: Correlation matrix of factor scores based on eating patterns at 6 years

	Occasional	**Seafood**	**Fruit and vegetables**	**Meat**
Occasional	1.000			
Seafood	0.309	1.000		
Fruit and vegetables	0.189	0.395	1.000	
Meat	0.408	0.421	0.418	1.000

Table 3.14: Correlation matrix of factor scores based on eating patterns at 9 years

	Occasional	**Seafood**	**Fruit and vegetables**	**Meat**
Occasional	1.000			
Seafood	0.277	1.000		
Fruit and vegetables	-0.185	0.044	1.000	
Meat	0.091	0.113	0.101	1.000

Table 3.15: Correlation matrix of factor scores based on eating patterns at 14 years

	Occasional	**Seafood**	**Fruit and vegetables**	**Meat**
Occasional	1.000			
Seafood	0.267	1.000		
Fruit and vegetables	0.026	0.208	1.000	
Meat	0.301	0.340	0.166	1.000

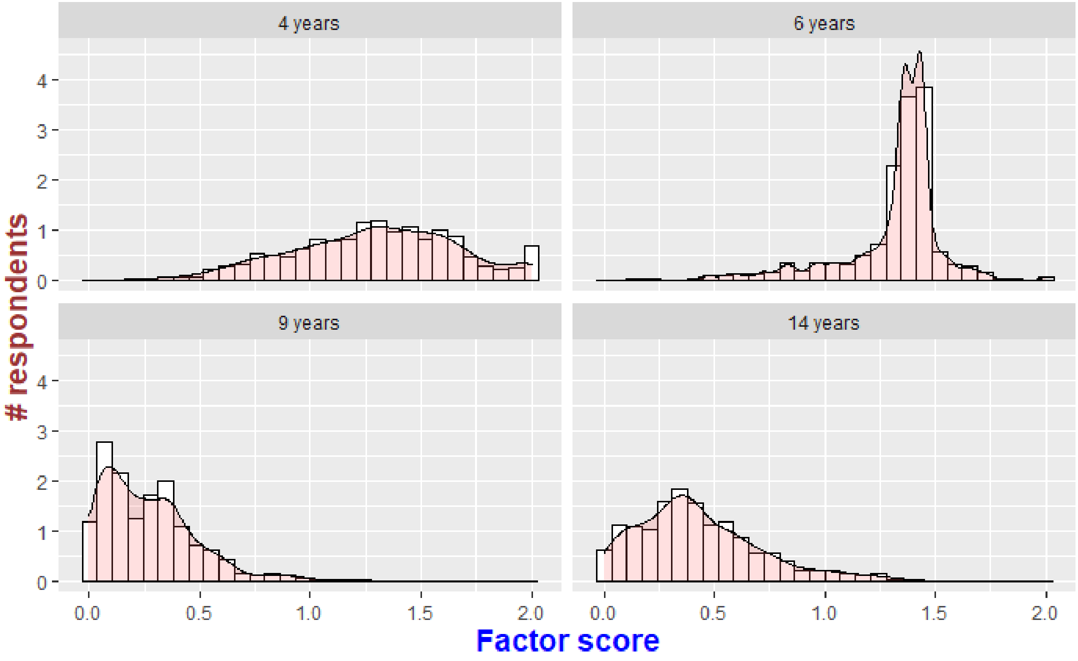
Smoothed densities of factor scores are displayed in Table 3.16 and Figure 3.2. The distribution of the ”Occasional” factor score is narrower at 6-years and wider at 4-years compared to the other ages. The scale of the ”Seafood” factor score is similar across all measurement waves. Most scores are located in the range from 0.0 to 0.5 portions per day at all ages. A spike occurs at value 4 in the ”Basics and staples” factor score at 4-years. Moreover, the distribution of ”Fruit and vegetables” factor score at 9 years and 14 years is notably different from the distribution at 4 years and 6 years, displaying approximate discreteness in the former. This is an effect of the differing dietary assessment methods across measurement waves and the computation of the ”Fruit and vegetables” score: consumption is reported directly in portions per day at 9 and 14 years, and only two food categories enter into the eating patterns with roughly equal weights of 0.54 and 0.46. In terms of scale, the ”Fruit and vegetables” factor score has a wide spread at all ages. The ”Meat” factor score is clearly more concentrated at 6 compared to 4 years. Its distribution is similar between 9-years and 14-years, in both of cases ranging approximately from 1 to 2. In general, because the factor scores are on the scale of daily portions, we expect a factor score with a narrow spread provides very little information when fitted as a linear regression covariate, compared to a score with a wider spread. Hence, we expect that factor scores at 6 years will not have a crucial impact in the causal analysis, since only 1 out of their 4 distributions is obviously scattered.

Table 3.16: Descriptive table of nutrition factor scores

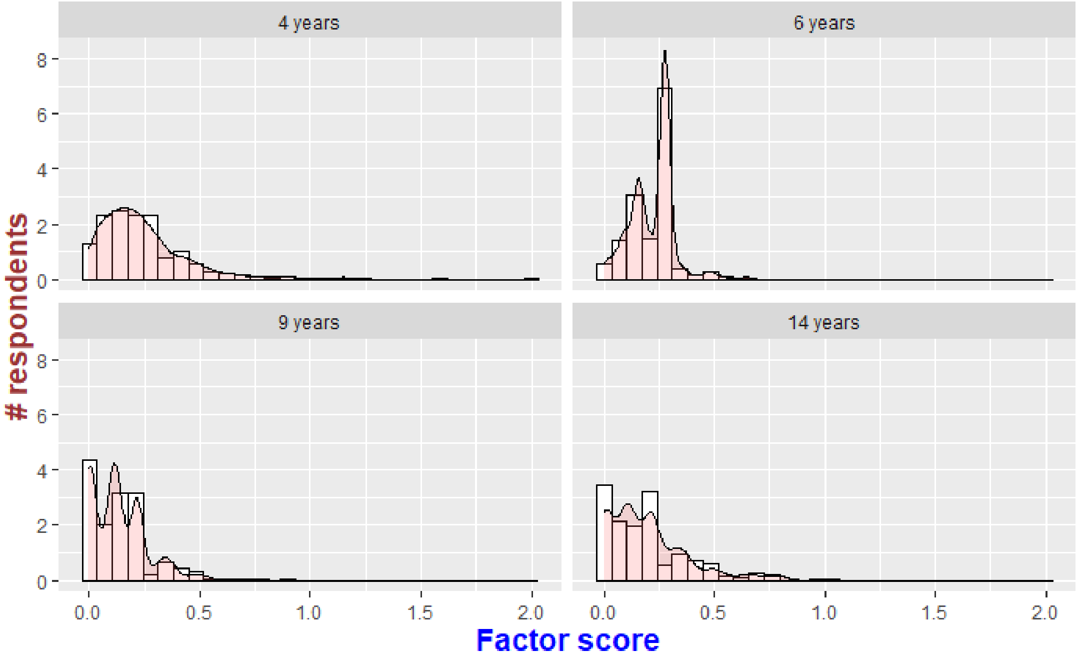
Eating pattern	n	miss	p.miss	mean	sd	median	p25	p75	min	max
4 years										
Occasional	1,398	491	35.1	1.0	0.4	1.0	1.0	2.0	0.2	2.0
Seafood	1,398	493	35.3	0.3	0.2	0.2	0.1	0.3	0.005	2.0
Fruit and vegetables	1,398	491	35.1	3.0	0.9	3.0	2.0	4.0	0.3	4.0
Meat	1,398	492	35.2	2.0	0.5	2.0	2.0	3.0	0.5	3.0
6 years										
Occasional	1,398	600	42.9	1.0	0.2	1.0	1.0	1.0	0.1	2.0
Seafood	1,398	599	42.8	0.2	0.1	0.3	0.2	0.3	0.007	1.0
Fruit and vegetables	1,398	597	42.7	2.0	0.7	3.0	2.0	3.0	0.1	4.0
Meat	1,398	597	42.7	2.0	0.3	2.0	2.0	2.0	0.5	3.0
9-years										
Occasional	1,398	423	30.3	0.3	0.2	0.2	0.1	0.4	0.0	1.0
Seafood	1,398	434	31.0	0.1	0.1	0.1	0.005	0.2	0.0	0.9
Fruit and vegetables	1,398	422	30.2	2.0	0.9	2.0	2.0	3.0	0.2	4.0
Meat	1,398	431	30.8	2.0	0.3	2.0	1.0	2.0	0.09	2.0
14 years										
Occasional	1,398	531	38.0	0.4	0.3	0.4	0.2	0.6	0.001	1.0
Seafood	1,398	945	67.6	0.2	0.2	0.1	0.09	0.3	0.005	1.0
Fruit and vegetables	1,398	566	40.5	2.0	1.0	2.0	1.0	3.0	0.0	4.0
Meat	1,398	818	58.5	2.0	0.3	2.0	1.0	2.0	0.8	2.0

Note: n - the total number of observations; miss - the number of observations with missing value; p.miss - the percentage of observations with missing value accounted for the total number; sd - standard deviation; p25 - 25% percentile; p75 - 75% percentile;

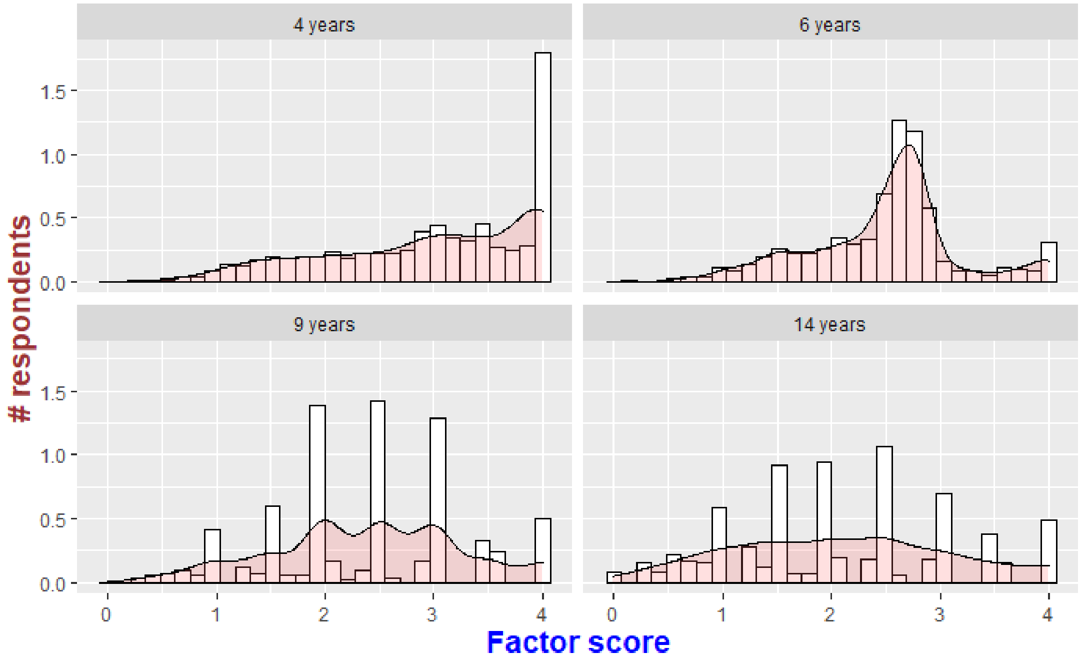
Figure 3.2: Density of factor score



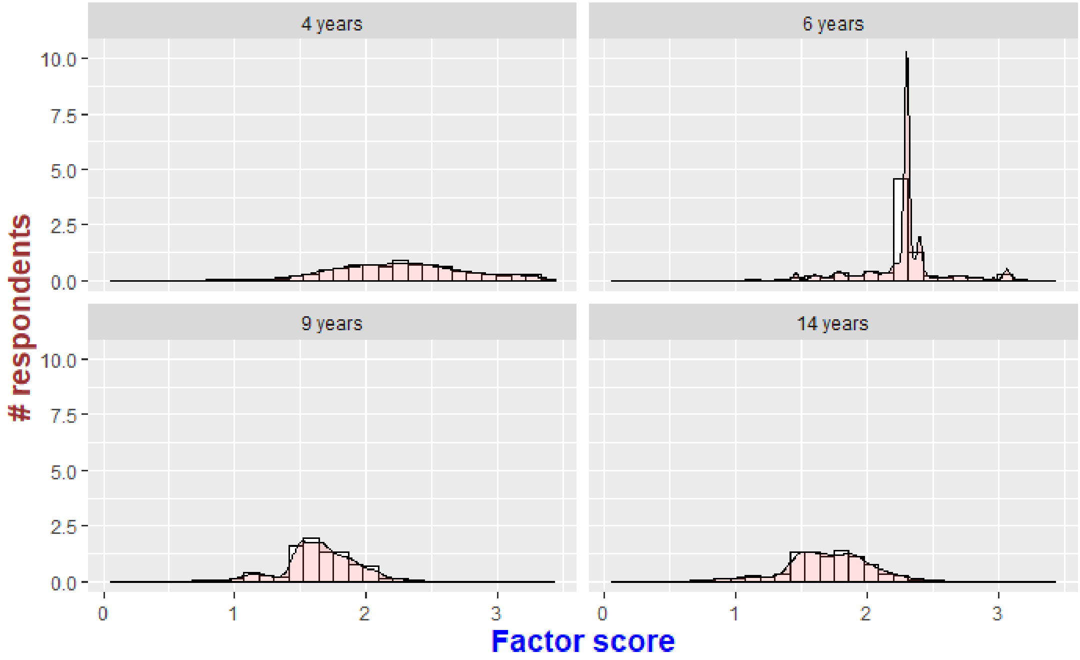
(a) Occasional



(b) Seafood



(c) Fruit and vegetables



(d) Meat

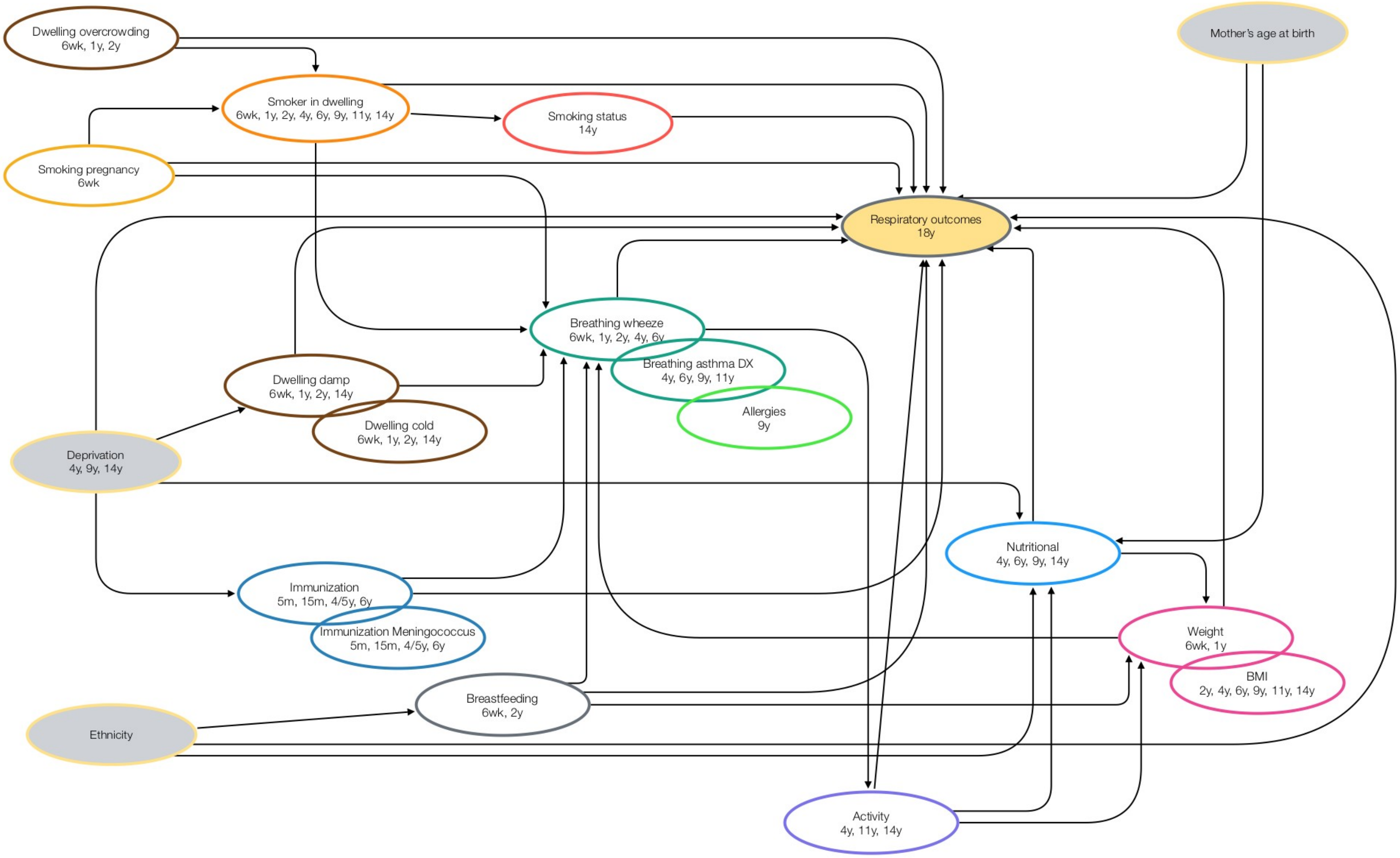
3.3 Causal diagram

Figure 3.3 is the summary causal diagram showing the causal paths between all exposures of interest and the final outcomes (Respiratory outcomes measured at 18 years); however it does not display the causal paths across different measurement waves. Figure 3.4, 3.5, 3.6, and 3.7 provide the detailed causal paths between nutrition factor scores and the final outcomes (FEV1 z-score measured at 18 years) at 4 years, 6 years, 9 years, and 14 years. These figures not only reveal the causal paths at the same measurement wave, but also across the different measurement waves. In the diagrams, we utilize different colours to identify the domain to which an exposure belongs. If exposures belong the same domain, they will be labelled with the same colour. The details of domain colour mapping are shown in Table 3.17.

Table 3.17: Domain colour mapping

Domain	**Colour**
General information	Yellow ellipse with grey background
Final outcome	Dark grey ellipse with orange background
Immunisation	Dark blue ellipse
Exercise	Purple ellipse
Breastfeeding	Grey ellipse
Antenatal smoking	Yellow ellipse
Smoking exposure	Orange ellipse
Smoking	Red ellipse
Respiratory illness-infection	Dark green ellipse
Weight, height, BMI	Pink ellipse
Nutrition	Light blue ellipse
Allergies	Light green ellipse
Dwelling	Brown ellipse

Figure 3.3: Summary causal diagram



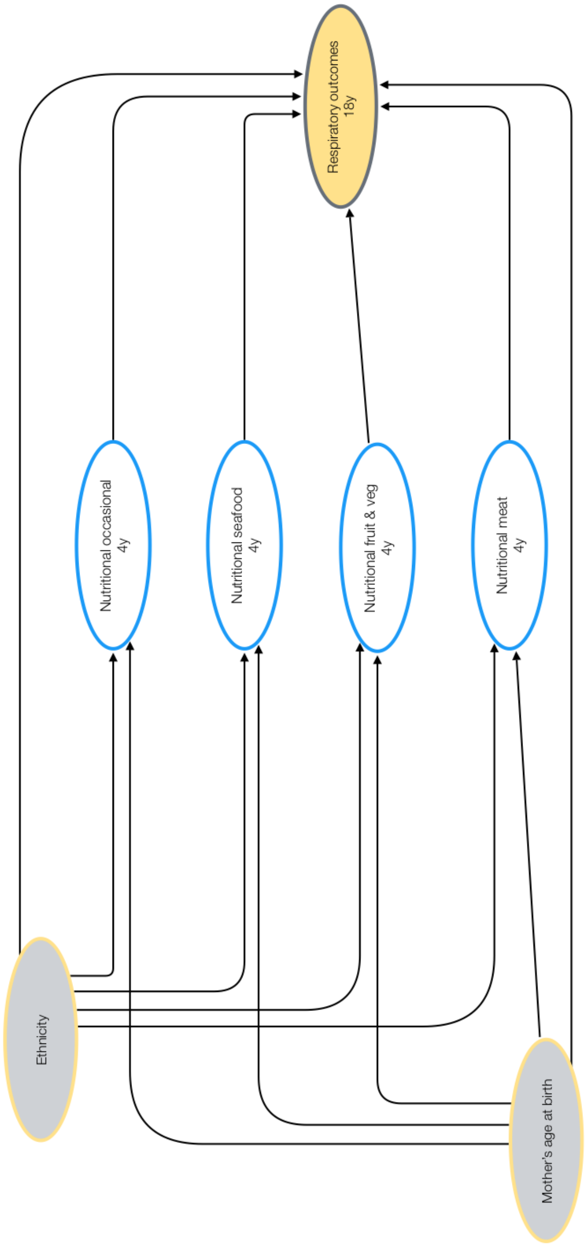


Figure 3.4: Causal diagram of nutrition factor scores at 4 years

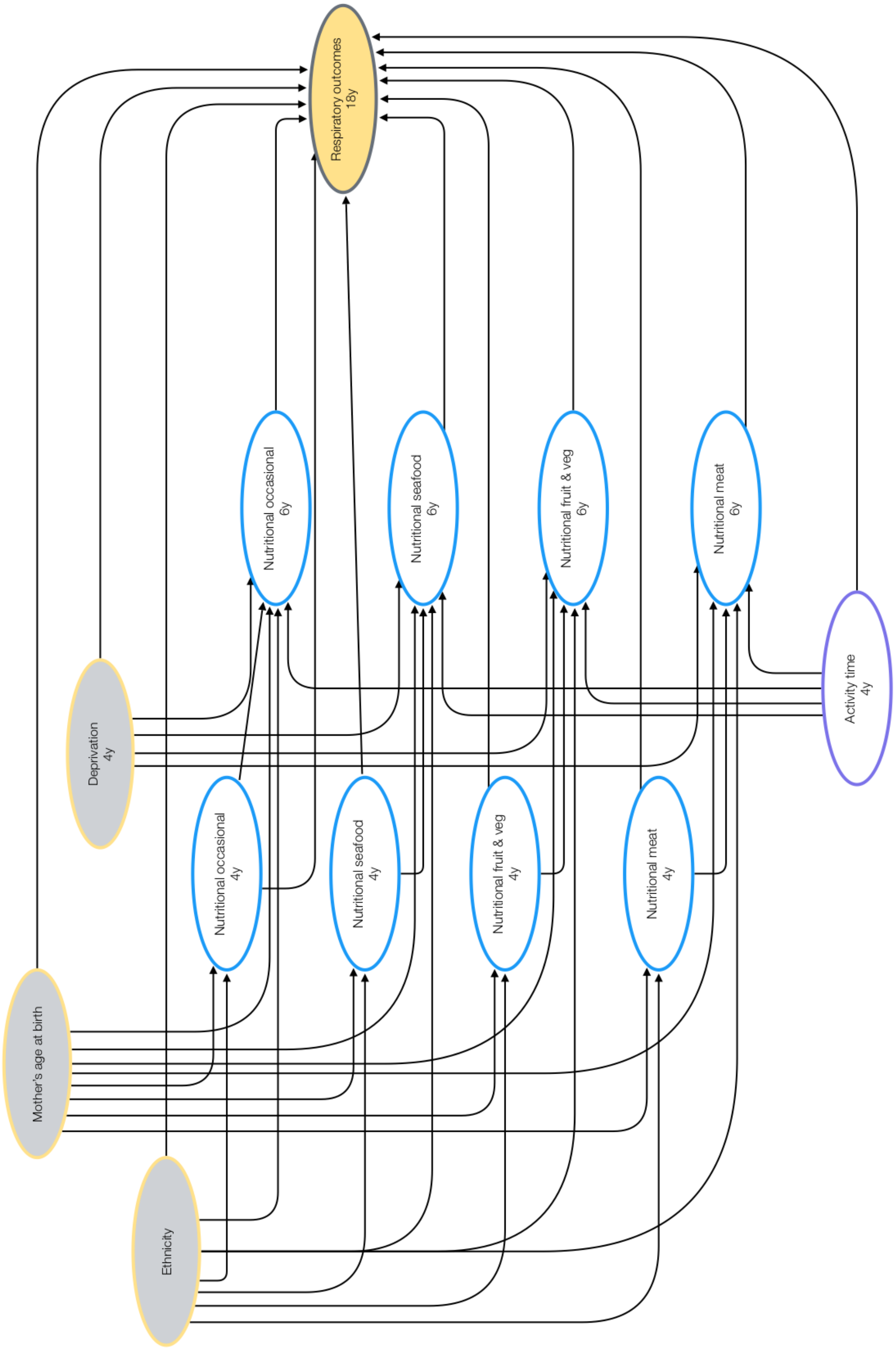


Figure 3.5: Causal diagram of nutrition factor scores at 6 years

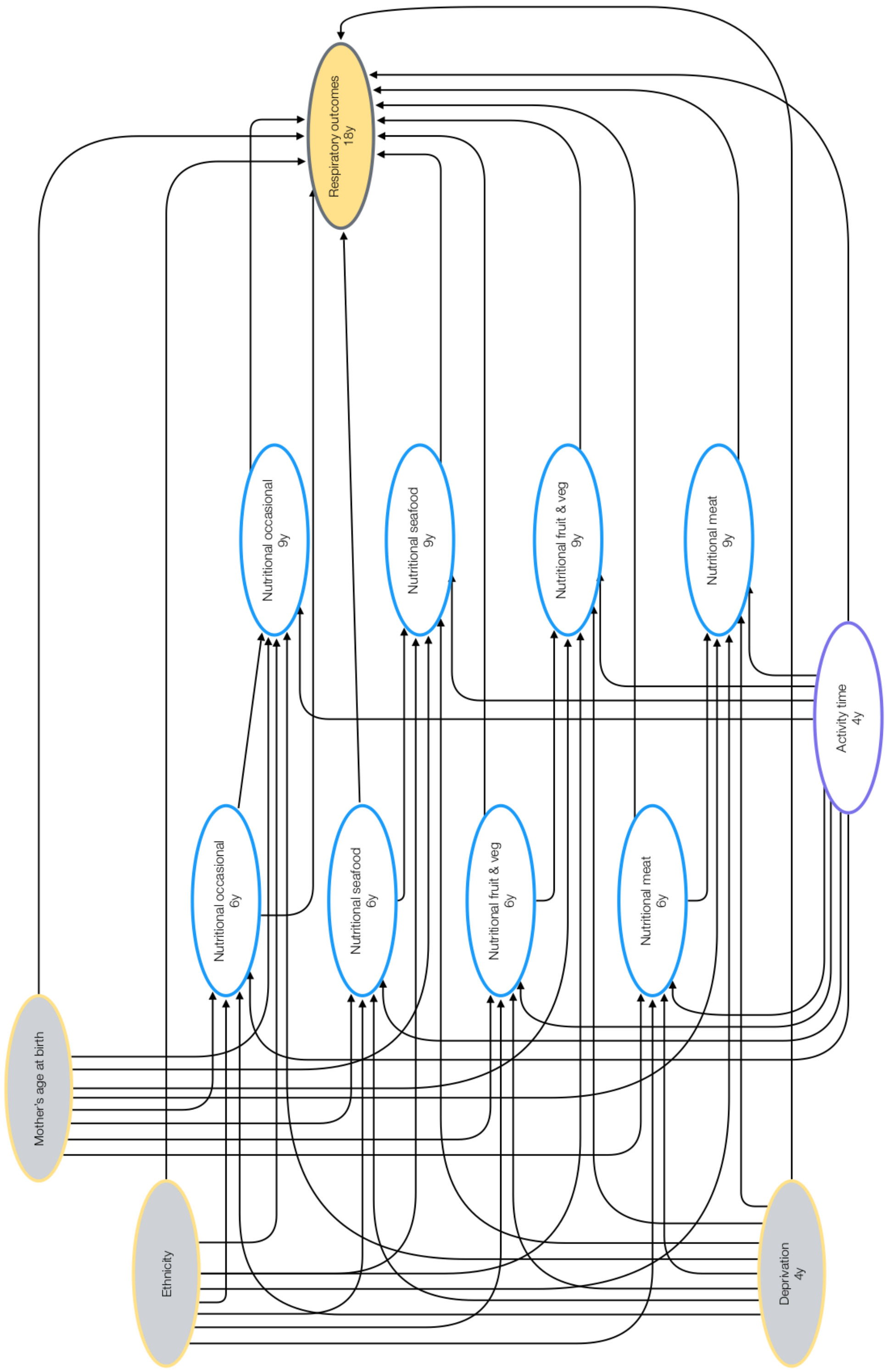


Figure 3.6: Causal diagram of nutrition factor scores at 9 years

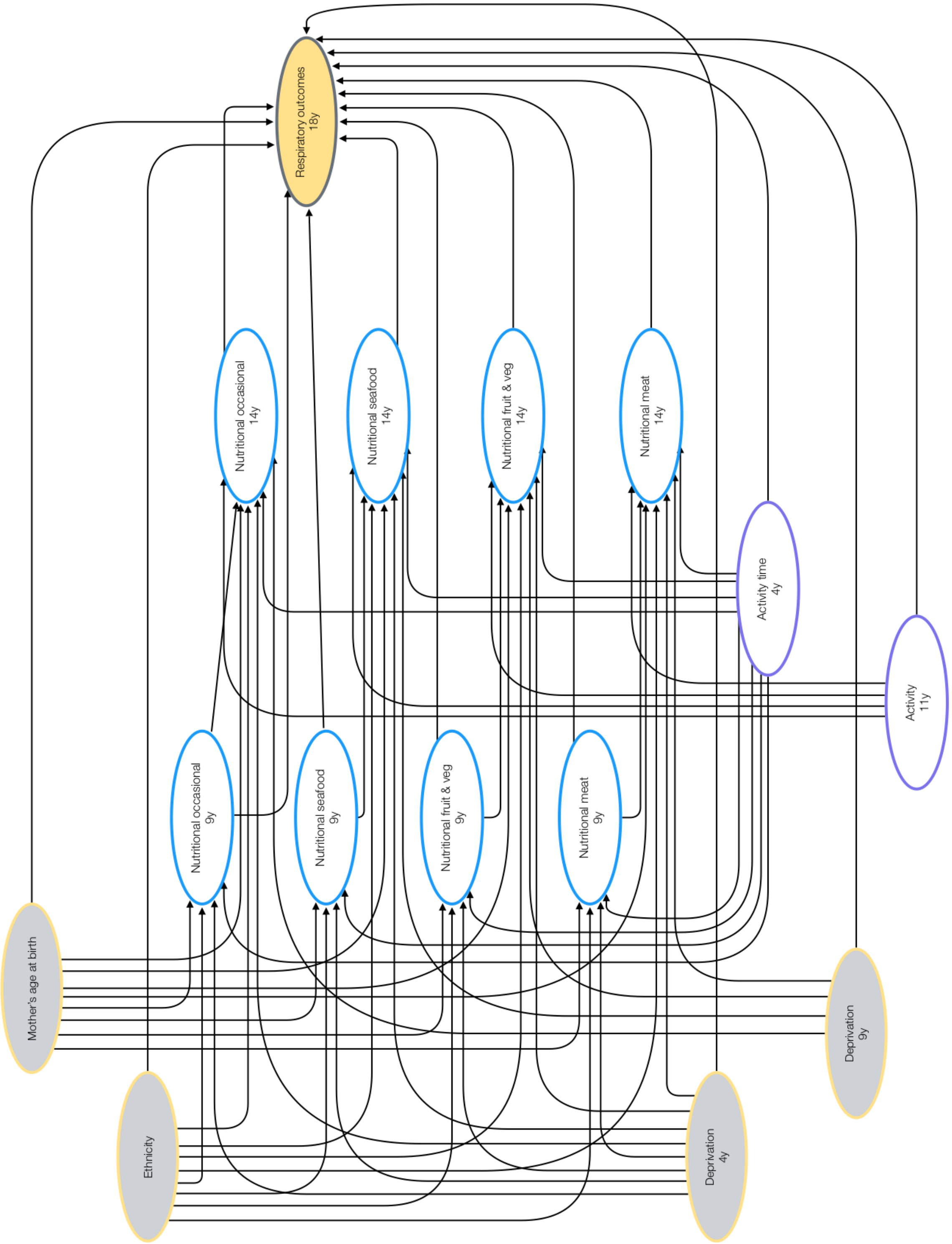


Figure 3.7: Causal diagram of nutrition factor scores at 14 years

3.4 Weight estimation

This section presents the results from the weighting models. Table 3.18 shows the baseline characteristics of all participants in the birth cohort, participants who did not attend 18-year measurement wave, and participants who did. As we can see, the distribution of some baseline characteristics are significantly different between the participants who did and did not attend 18-year measurement wave, such as: mother's highest school qualification; mother's highest post-school qualification; whether mother was employed prior to pregnancy, etc. Therefore, the PIFS cohort data at 18-years indeed contains selection bias and needs to be re-balanced by weights to reduce the impact from selection bias. Table 3.19 shows the proportion of missing values in exposures and confounders by measurement wave and eating pattern. We observed that the proportion of missing values in exposures and confounders can differ within the same eating pattern across different measurement waves, or within the same measurement wave across different eating pattern. It was therefore preferable to apply an independent model of missingness for each eating pattern in each measurement wave to calculate the relevant weights. This is what we did in this thesis.

Table 3.18: Baseline characteristics amongst birth cohort and participants who did and di not attend the 18-year measurement wave

Domain	**Birth cohort**	**Not attend 18-year wave**	**Attend 18-year wave**	**p-value**
General Information				
Parental Ethnicity				
Samoan				0.3150
Both of parents	541 (38.7%)	171 (36.7%)	370 (39.7%)	
One of parents	219 (15.7%)	69 (14.8%)	150 (16.1%)	
Neither parent	638(45.6%)	226 (48.5%)	412 (44.2%)	
Cook Island				0.2030
Both of parents	117 (8.4%)	40 (8.6%)	77 (8.3%)	
One of parents	172 (12.3%)	47 (10.1%)	125 (13.4%)	
Neither parent	1,109 (79.3%)	379 (81.3%)	730 (78.3%)	
Tongan				0.0020
Both of parents	260 (18.6%)	109 (23.4%)	151 (16.2%)	
One of parents	113 (8.1%)	29 (6.2%)	84 (9.0%)	
Neither parent	1,025 (73.3%)	328 (70.4%)	697 (74.8%)	

Continued on next page

Table 3.18 – *Continued from previous page*

Domain	**Birth cohort**	**Did not attend 18-year wave**	**Attended 18-year wave**	**p-value**
Other: Pacific Island				0.0260
Both of parents	36 (2.6%)	17 (3.6%)	19 (2.0%)	
One of parents	135 (9.7%)	34 (7.3%)	101 (10.8%)	
Neither parent	1,227 (87.8%)	415 (89.1%)	812 (87.1%)	
Other: Non Pacific Island				0.5880
Both of parents	18 (1.3%)	4 (0.9%)	14 (1.5%)	
One of parents	202 (14.4%)	69 (14.8%)	133 (14.3%)	
Neither parent	1,178 (84.3%)	393 (84.3%)	785 (84.2%)	
Sex				
Baby sex				0.0030
Male	717 (51.3%)	505 (54.2%)	212 (45.5%)	
Female	681 (48.7%)	427 (45.8%)	254 (54.5%)	
Living environment				
With both natural parents				0.4930
Yes	1,061 (75.9%)	713 (76.5%)	348 (74.7%)	
No	337 (24.1%)	219 (23.5%)	118 (25.3%)	
With adoptive parents				1.0000
Yes	114 (8.2%)	76 (8.2%)	38 (8.2%)	
No	1,284 (91.8%)	856 (91.8%)	428 (91.8%)	
With single parent family				0.9250
Yes	141 (10.1%)	93 (10.0%)	48 (10.3%)	
No	1,257 (89.9%)	839 (90.0%)	418 (89.7%)	
With a step parent				0.0740
Yes	25 (1.8%)	12 (1.3%)	13 (2.8%)	
No	1,373 (98.2%)	920 (98.7%)	453 (97.2%)	
In another relative's home				0.7480
Yes	166 (11.9%)	113 (12.1%)	53 (11.4%)	
No	1,232 (88.1%)	819 (87.9%)	413 (88.6%)	
In a foster family				0.5400
Yes	3 (0.3%)	3 (0.3%)	0 (0.0%)	

Continued on next page

Table 3.18 – *Continued from previous page*

Domain	**Birth cohort**	**Did not attend 18-year wave**	**Attended 18-year wave**	**p-value**
No	1,395 (99.8%)	929 (99.7%)	466 (100.0%)	
In a welfare family				1.0000
Yes	1 (0.1%)	1 (0.1%)	0 (0.0%)	
No	1,397 (99.9%)	931 (99.9%)	466 (100.0%)	
In an institution				NA
Yes				
No	1,398 (100.0%)	932 (100.0%)	466 (100.0%)	
Other				0.9290
Yes	16 (1.1%)	10 (1.1%)	6 (1.3%)	
No	1,382 (98.9%)	922 (98.9%)	460 (98.7%)	
Education				
Mother's highest school qualification				0.0100
No formal qualification	595 (42.6%)	422 (45.3%)	173 (37.1%)	
NZ 6th Form Cert. in 1 or more subjects	177 (12.7%)	105 (11.3%)	72 (15.5%)	
NZ Higher School Cert. or Higher Leaving Cert.	29 (2.1%)	14 (1.5%)	15 (3.2%)	
NZ School Cert. in 1 or more subjects	452 (32.3%)	298 (32.0%)	154 (33.0%)	
NZ U.E. (pre '86) in 1 or more subjects	45 (3.2%)	25 (2.7%)	20 (4.3%)	
NZ University Bursary or Scholarship	60 (4.3%)	38 (4.1%)	22 (4.7%)	
Overseas secondary school qualification	32 (2.3%)	23 (2.5%)	9 (1.9%)	
Other NZ secondary school qualification	8 (0.6%)	7 (0.8%)	1 (0.2%)	
Mother's highest post-school qualification				0.0100
Advanced Trade Certificate	30 (2.1%)	20 (2.1%)	10 (2.1%)	
Bachelors Degree	18 (1.3%)	11 (1.2%)	7 (1.5%)	

Continued on next page

Table 3.18 – *Continued from previous page*

Domain	**Birth cohort**	**Did not attend 18-year wave**	**Attended 18-year wave**	**p-value**
No other qualifications	1,012 (72.4%)	686 (73.6%)	326 (70.0%)	
NZ Certificate or Diploma	66 (4.7%)	40 (4.3%)	26 (5.6%)	
Polytechnic Certificate or Diploma	88 (6.3%)	54 (5.8%)	34 (7.3%)	
Post-graduate Degree, Certificate or Diploma	2 (0.1%)	2 (0.2%)	0 (0.0%)	
Teachers Certificate or Diploma	32 (2.3%)	18 (1.9%)	14 (3.0%)	
Technicians Certificate	7 (0.5%)	3 (0.3%)	4 (0.9%)	
Trade Certificate	129 (9.2%)	89 (9.5%)	40 (8.6%)	
Employment				
Mother's employment situation prior to pregnancy				0.0510
Full-time paid workforce (1 job)	582 (41.6%)	361 (38.7%)	221 (47.4%)	
Full-time paid workforce (2 or more jobs)	10 (0.7%)	8 (0.9%)	2 (0.4%)	
Full-time parent (unpaid)	312 (22.3%)	220 (23.6%)	92 (19.7%)	
Part-time paid workforce (1 job)	165 (11.8%)	107 (11.5%)	58 (12.4%)	
Part-time paid workforce (2 or more jobs)	6 (0.4%)	3 (0.3%)	3 (0.6%)	
Student	66 (4.7%)	176 (18.9%)	70 (15.0%)	
Other	11 (0.8%)	8 (0.9%)	3 (0.6%)	
Mother employed prior to pregnancy				0.0001
Yes	194 (13.9%)	108 (11.6%)	86 (18.5%)	
No	1,204 (86.1%)	824 (88.4%)	380 (81.5%)	
Mother's employment				0.8080

Continued on next page

Table 3.18 – *Continued from previous page*

Domain	**Birth cohort**	**Did not attend 18-year wave**	**Attended 18-year wave**	**p-value**
Employer paid parental leave	7 (0.5%)	5 (0.5%)	2 (0.4%)	
Full time work	56 (4.0%)	37 (4.0%)	19 (4.1%)	
Not in regular employment	1,307 (93.5%)	869 (93.2%)	438 (94.0%)	
Part time work	28 (2.0%)	21 (2.3%)	7 (1.5%)	
Income				
Household income				0.1530
$0 - $20,000	466 (33.3%)	316 (33.9%)	150 (32.2%)	
$20,001 - $40,000	717 (51.3%)	471 (50.5%)	246 (52.8%)	
> $40,000	165 (11.8%)	105 (11.3%)	60 (12.9%)	
Unknown	50 (3.6%)	40 (4.3%)	10 (2.1%)	
Age				
Mother's birth age				< 0.0010
< 20	117 (7.9%)	90 (9.7%)	21 (4.5%)	
20 - 29	733 (52.4%)	496 (53.2%)	237 (50.9%)	
30 - 39	508 (36.3%)	325 (34.9%)	183 (39.3%)	
40+	46 (3.3%)	21 (2.3%)	25 (5.4%)	
Immunisation				
Immunisation				
Immunized at 6 weeks				0.4690
Immunized at 6 weeks	mean: 0.73 (sd: 0.44)	mean: 0.74 (sd: 0.44)	mean: 0.72 (sd: 0.45)	
Breast feeding				
Breast feeding				
How fed in 1st 6 weeks				0.9100
Combination breast milk and other	532 (38.1%)	356 (38.2%)	176 (37.8%)	
Only with breast milk	686 (49.1%)	454 (48.7%)	232 (49.8%)	
Only with formula or other bottle milk	180 (12.9%)	122 (13.1%)	58 (12.4%)	

Continued on next page

Table 3.18 – *Continued from previous page*

Domain	**Birth cohort**	**Did not attend 18-year wave**	**Attended 18-year wave**	**p-value**
Antenatal smoking				
Antenatal smoking				
Smoked during pregnancy				0.0560
Yes	345 (24.7%)	245 (26.3%)	100 (21.5%)	
No	1,053 (75.3%)	687 (73.7%)	366 (78.5%)	
# cigarettes during pregnancy				0.1270
# cigarettes during pregnancy	mean: 0.07 (sd: 0.17)	mean: 0.07 (sd: 0.17)	mean: 0.06 (sd: 0.15)	
Smoking exposure				
Smoking exposure				
Smoking people living in house hold now				0.5080
Yes	716 (51.2%)	471 (50.5%)	245 (52.6%)	
No	682 (48.8%)	461 (49.5%)	221 (47.4%)	
Respiratory illness-infection				
Respiratory illness-infection				
Problems with breathing				0.4330
Yes	598 (42.8%)	406 (43.6%)	192 (41.2%)	
No	800 (57.2%)	526 (56.4%)	274 (58.8%)	
Weight, height, BMI				
Weight, height, BMI				
Birth weight				0.8250
Birth weight	mean: 3,568.69 (sd: 622.16)	mean: 3,571.30 (sd: 624.43)	mean: 3,563.49 (sd: 618.21)	
Dwelling				
Dwelling				
Dampness/mould				0.0690
Yes	522 (37.3%)	332 (35.6%)	190 (40.8%)	
No	876 (62.7%)	600 (64.4%)	276 (59.2%)	
Cold				0.8640
Yes	753 (53.9%)	504 (54.1%)	249 (53.4%)	

Continued on next page

Table 3.18 – *Continued from previous page*

Domain	**Birth cohort**	**Did not attend 18-year wave**	**Attended 18-year wave**	**p-value**
No	645 (46.1%)	428 (45.9%)	217 (46.6%)	
Overcrowding				0.7280
Yes	431 (30.8%)	284 (30.5%)	147 (31.5%)	
No	967 (69.2%)	648 (69.5%)	319 (68.5%)	

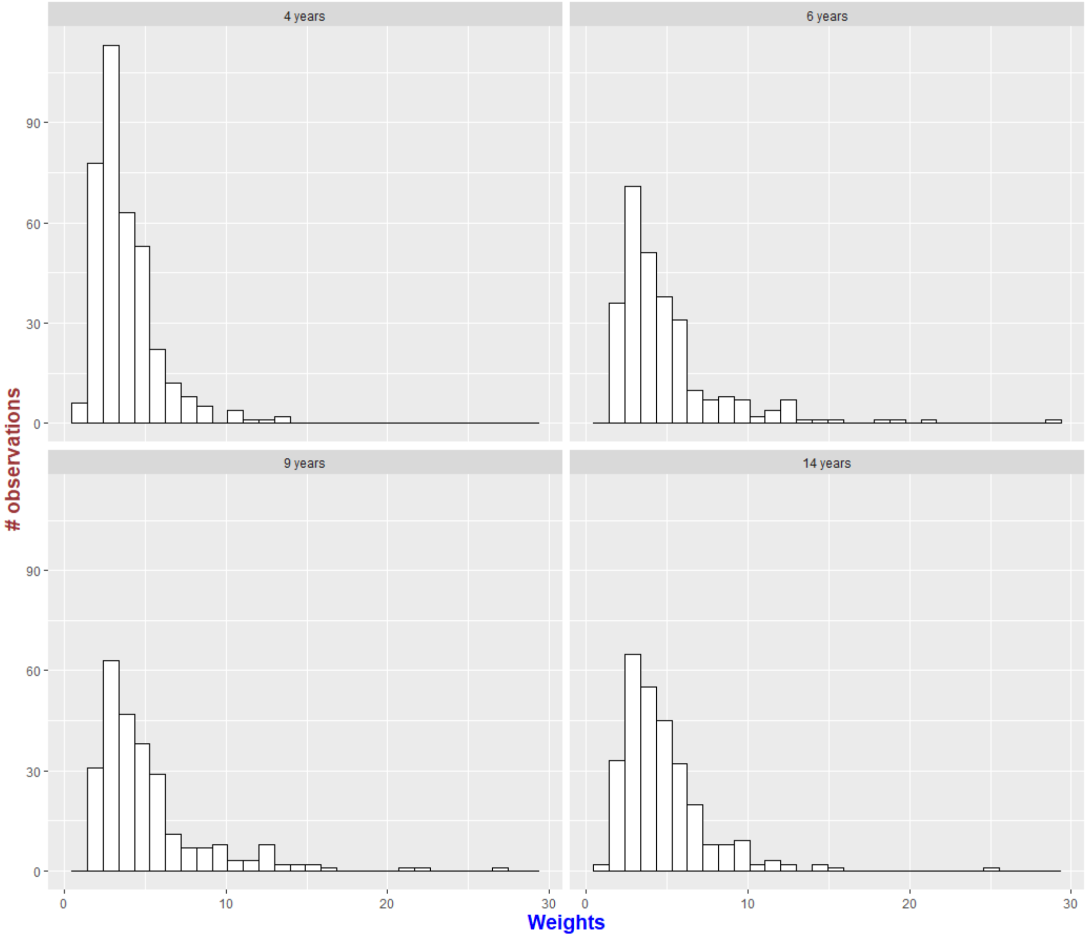
Note: The p-value is computed from the comparison between not attend 18-year measurement wave and attend 18-year measurement wave

Table 3.19: The proportion of missing values in exposures and confounders by measurement waves and eating patterns

Eating pattern	**Exposure missing**	**Confounder missing**
4 years		
Occasional	98 (21%)	NA
Seafood	98 (21%)	NA
Fruit and vegetables	98 (21%)	NA
Meat	98 (21%)	NA
6 years		
Occasional	126 (27%)	50 (15%)
Seafood	127 (27%)	49 (14%)
Fruit and vegetables	125 (27%)	50 (15%)
Meat	125 (27%)	50 (15%)
9-years		
Occasional	35 (8%)	107 (25%)
Seafood	40 (9%)	106 (25%)
Fruit and vegetables	34 (7%)	106 (25%)
Meat	37 (8%)	105 (24%)
14 years		
Occasional	44 (9%)	25 (6%)
Seafood	247 (53%)	16 (7%)
Fruit and vegetables	62 (13%)	24 (6%)
Meat	194 (12%)	19 (7%)

Figure 3.8 shows the distribution of inverse probability weights (IPW) for the ”Fruit and vegetables” eating pattern across measurement waves. The distribution of weights for other eating patterns (Figure A.1, A.2, A.3) are similar.

Figure 3.8: The distribution of inverse probability weights (IPW) for the ”Fruit and vegetables” eating pattern across measurement waves



3.5 Causal results

3.5.1 Linear regression model

Figure 3.20, 3.21, and 3.22 respectively show the estimated causal effects of nutrition factor scores on FEV1 adjusted for height and sex, FEV1 Z-score, and FEV1 % predicted. These causal results are obtained from the weighted linear regression models adjusted for confounders, as determined from the causal diagram. As we can see, the causal results for FEV1 Z-score and FEV1 % predicted have the same signs in each eating pattern. This means that, as expected, in each eating pattern, nutrition factor scores impose the causal effects on FEV1 Z-score and FEV1 % predicted in the same direction. The linear regression models estimate that the causal effect of most of the eating patterns on FEV1 Z-score and FEV1 % predicted is negative at 4 years as well as at 14 years, but is positive at 6 years and at 9 years. Another point to note is that the estimated causal effect of ”Seafood” eating pattern on FEV1 Z-score and FEV1 % predicted is only positive at the 6-year measurement wave and negative at other measurement waves. However, the estimates for other eating patterns are positive at most of the measurement waves. Nevertheless, nutrition factor scores impose the effects on FEV1 adjusted for height and sex in the opposite direction in the ”Meat” eating patterns at 6 years and at 14 years as well as in the ”Seafood” eating pattern at 14 years when compared to the other results.

The pattern of standard errors is similar in all causal results. The standard errors of the estimated causal effects of nutrition factor scores on FEV1 Z-score, FEV1 % predicted, and FEV1 adjusted for height and sex at 4-years measurement wave is remarkably smaller than most of the eating patterns in other measurement waves. Furthermore, when comparing the spread of the estimates in all causal results from the eating patterns perspective, we notice that estimate variability is smallest in the ”Fruit and vegetables” eating pattern, and largest in the ”Seafood” eating pattern. In relative terms, the standard errors are relatively close in the remaining two eating patterns.

FEV1 Z-score, FEV1 % predicted, and FEV1 adjusted for height and sex respectively display two causal results with significant p-value. The significant estimated causal effects of nutrition factor scores on FEV1 Z-score and FEV1 % predicted occur in the ”Fruit and vegetables” eating pattern at 4 years (Original range in portions per day: (0.3, 4.0)) and 9 years (Original range in portions per day: (0.2, 4.0)). We estimate that, on average, one added portion per day of ”Fruit and vegetables” will increase FEV Z-score by 0.13 units at 4 years and 0.25 units at 9 years, and will increase FEV1 % predicted by 1.59 percentage points at 4 years and 2.94 percentage points at 9 years. The significant causal results of FEV1 adjusted for height and sex occur in the ”Fruit and vegetables” eating pattern (Original range in portions per day: (0.3, 4.0)) and the ”Meat” eating pattern at 9 years (Original range in portions per day: (0.09, 2.0)). We estimate that, on average, one added portion per day of ”Fruit and vegetables” will increase lung function by 120 millilitres (95% CI (0, 210)), while one added portion per day in the ”Meat” eating pattern will increase lung function by 290 millilitres (95% CI (-20, 570)) at 9 years.

Table 3.20: The causal effects of nutrition factor scores on FEV1 adjusted for height and sex (Linear regression)

Eating pattern	**Estimate** (liter)	**Standard Error**	**95% Confidence Interval**	**p-value**
4 years				
Occasional	-0.07	0.08	(-0.21, 0.09)	0.5621
Seafood	-0.05	0.11	(-0.26, 0.16)	0.6964
Fruit and vegetables	0.04	0.04	(-0.03, 0.11)	0.1998
Meat	-0.04	0.05	(-0.14, 0.06)	0.8586
6 years				
Occasional	0.21	0.19	(-0.16, 0.57)	0.3596
Seafood	0.22	0.29	(-0.40, 0.77)	0.5033
Fruit and vegetables	0.05	0.06	(-0.07, 0.15)	0.5996
Meat	-0.05	0.11	(-0.28, 0.18)	0.7204
9-years				
Occasional	0.21	0.19	(-0.15, 0.57)	0.2632
Seafood	-0.16	0.39	(-0.99, 0.56)	0.8280
Fruit and vegetables	0.12	0.05	(0.00, 0.21)	0.0139
Meat	0.29	0.16	(-0.02, 0.57)	0.0700
14 years				
Occasional	0.07	0.13	(-0.19, 0.33)	0.7061
Seafood	0.09	0.27	(-0.42, 0.65)	0.7184
Fruit and vegetables	-0.04	0.03	(-0.10, 0.03)	0.3097
Meat	0.00	0.16	(-0.36, 0.26)	0.9372

Table 3.21: The causal effects of nutrition factor scores on FEV1 Z-score (Linear regression)

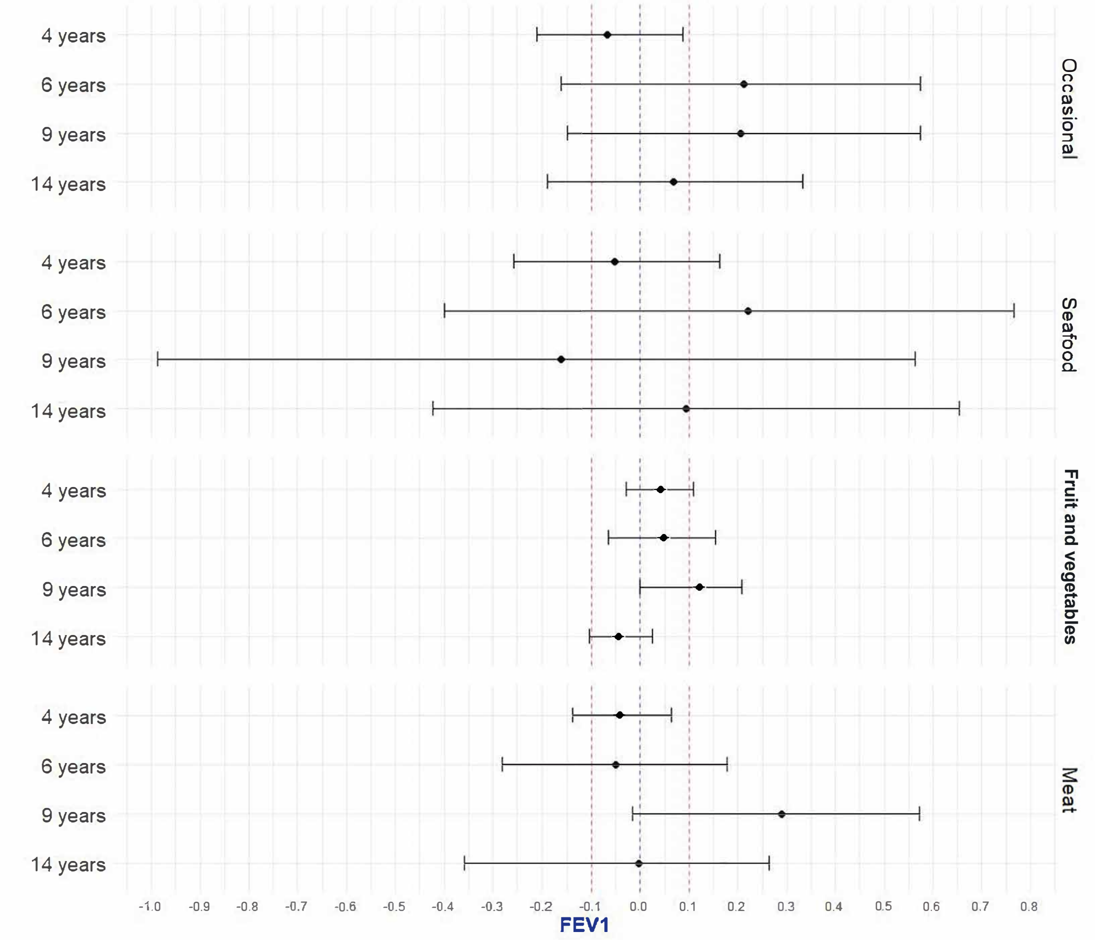
Eating pattern	**Estimate**	**Standard Error**	**95% Confidence Interval**	**p-value**
4 years				
Occasional	-0.10	0.16	(-0.40, 0.21)	0.6562
Seafood	-0.06	0.22	(-0.49, 0.38)	0.7673
Fruit and vegetables	0.13	0.08	(-0.01, 0.27)	0.0790
Meat	-0.07	0.11	(-0.27, 0.15)	0.9088
6 years				
Occasional	0.44	0.40	(-0.35, 1.19)	0.4216
Seafood	0.70	0.68	(-0.86, 1.82)	0.3667
Fruit and vegetables	0.17	0.12	(-0.08, 0.39)	0.2379
Meat	0.03	0.27	(-0.52, 0.56)	0.8653
9-years				
Occasional	0.26	0.38	(-0.46, 1.04)	0.5023
Seafood	-0.10	0.92	(-2.01, 1.63)	0.9133
Fruit and vegetables	0.25	0.11	(0.00, 0.43)	0.0170
Meat	0.50	0.31	(-0.11, 1.07)	0.2105
14 years				
Occasional	0.09	0.29	(-0.49, 0.69)	0.7231
Seafood	-0.06	0.56	(-1.08, 1.14)	0.8834
Fruit and vegetables	-0.10	0.07	(-0.22, 0.04)	0.3224
Meat	0.09	0.32	(-0.64, 0.65)	0.9037

Table 3.22: The causal effects of nutrition factor scores on FEV1 % Predicted (Linear regression)

Eating pattern	**Estimate** (p.p.)	**Standard Error**	**95% Confidence Interval**	**p-value**
4 years				
Occasional	-1.19	1.85	(-4.70, 2.56)	0.6723
Seafood	-0.80	2.61	(-5.77, 4.43)	0.7507
Fruit and vegetables	1.59	0.89	(-0.12, 3.20)	0.0842
Meat	-0.87	1.24	(-3.23, 1.70)	0.9105
6 years				
Occasional	5.15	4.64	(-4.16, 13.95)	0.4267
Seafood	8.04	8.02	(-10.42, 21.19)	0.3823
Fruit and vegetables	2.01	1.47	(-0.99, 4.57)	0.2447
Meat	0.29	3.13	(-6.18, 6.46)	0.8902
9-years				
Occasional	3.12	4.46	(-5.38, 12.23)	0.4941
Seafood	-1.59	10.88	(-24.17, 18.85)	0.8983
Fruit and vegetables	2.94	1.29	(0.00, 4.99)	0.0180
Meat	5.76	3.70	(-1.37, 12.59)	0.2138
14 years				
Occasional	1.01	3.49	(-5.88, 8.03)	0.7436
Seafood	-0.85	6.55	(-12.78, 13.17)	0.8660
Fruit and vegetables	-1.20	0.83	(-2.66, 0.47)	0.3131
Meat	0.98	3.81	(-7.60, 7.54)	0.9109

Note: p.p. - percentage point

Figure 3.9: The causal effects of nutrition factor scores on FEV1 adjusted for height and sex



FEV1 Z-score and FEV1 % Predicted are adequate to detect the statistical significance of differences, but they are not able to be used for identifying clinical significance. Compared to FEV1 Z-score and FEV1 % Predicted, FEV1 adjusted for height and sex is more suitable for identifying and interpreting clinical significance. Based on James F Donohue's paper [27], if the estimated causal effect of nutrition scores can make a change of 100mL or above in FEV1, it will be considered as clinically significant. To better identify which results are clinically significant, we utilise the forest plot to visualize the estimated causal effects of nutrition scores on FEV1 adjusted for height and sex. As Figure 3.9 shows, when comparing the range of 95% confidence interval over all eating patterns, the "Seafood" eating pattern has the widest range while the "Fruit and vegetables" eating pattern has the narrowest range. This is consistent with the standard errors given by the related linear regression model. In terms of statistical

significance and clinical significance, we can classify the causal results of FEV1 adjusted for height and sex to 4 categories:

- Statistically significance and clinically significance: The ”Fruit and vegetables” eating pattern at 9 years.
- Only statistically significance: None of the eating patterns.
- Only clinically significance: The ”Occasional” eating pattern at 6 years and 9 years, The ”Seafood” eating pattern at 6 years and 9 years, The ”Meat” eating pattern at 9 years. The 95% confidence interval for the ”Seafood” pattern are consistent with a clinically significant effect in either direction; the ”Meat” eating pattern at 9 years may warrant further investigation.
- Neither statistically significance nor clinically significance: The ”Occasional” eating pattern at 4 years and 14 years, The ”Seafood” eating pattern at 4 years and 14 years, The ”Fruit and vegetables” eating pattern at 4 years, 6 years, and 14 years, The ”Meat” eating pattern at 4 years, 6 years, and 14 years.

3.6 The results of population attributable fraction

3.6.1 Relative risk model

The results of relative risk models are shown in Table 3.23. They tell us that the relative risks of nutrition factor scores on the indicator of beneficial outcome are positive in most of the eating patterns. This means that, in most of the eating patterns, higher nutrition factor scores are protective for lung function and can lower the risk of obtaining respiratory disease in adulthood. However, there are some exceptions. The relative risks of the ”Meat” eating pattern are negative at 6 years and 14 years. This hints that higher nutrition factor scores in the ”Meat” eating pattern at 6 years and 14 years have detrimental effects on later lung function. As seen in 3.23, the estimated log-relative risks of nutrition factor scores on the indicator of beneficial outcome support some of the smallest standard errors at 4 years measurement wave in each eating pattern. From the point of view of the eating patterns, the standard errors are smallest in the ”Fruit and vegetables” eating pattern and largest in the ”Seafood” eating pattern. In the relative risk models, there are significant p-value in the ”Seafood” eating pattern at 4-years measurement wave and the ”Occasional” eating pattern at 9-years measurement wave. We estimate that, on average, for each added portion per day of ”Seafood”, the risk of healthy lung function will increase by 8.19% (Original range in portions per day of the ”Seafood” eating pattern at 4 years: (0.005, 2.0) Table: 3.16), and for each added portion per day of ”Occasional”, the risk of healthy lung function will increase by 11.30% (Original range in portions per day of the ”Occasional” eating pattern at 9 years: (0.0, 1.0) Table: 3.16).

Table 3.23: The causal effects of nutrition factor scores on the risk indicator extracted from FEV1 Z-score (Relative Risk)

Eating pattern	Estimate	Standard Error	95% Confidence Interval	p-value
4 years				
Occasional	0.04	0.04	(-0.02, 0.11)	0.1993
Seafood	0.08	0.04	(-0.01, 0.14)	0.0113
Fruit and vegetables	0.02	0.02	(-0.01, 0.05)	0.7403
Meat	0.04	0.03	(-0.01, 0.09)	0.3666
6 years				
Occasional	0.11	0.09	(-0.09, 0.27)	0.3822
Seafood	0.17	0.12	(-0.12, 0.36)	0.2054
Fruit and vegetables	0.00	0.03	(-0.05, 0.05)	0.8923
Meat	-0.05	0.06	(-0.14, 0.10)	0.4459
9-years				
Occasional	0.11	0.07	(-0.02, 0.24)	0.0453
Seafood	0.02	0.13	(-0.22, 0.32)	0.7965
Fruit and vegetables	0.05	0.02	(0.00, 0.08)	0.2025
Meat	0.04	0.06	(-0.08, 0.17)	0.5050
14 years				
Occasional	0.02	0.05	(-0.11, 0.09)	0.7682
Seafood	0.05	0.12	(-0.24, 0.24)	0.6905
Fruit and vegetables	0.00	0.01	(-0.03, 0.02)	0.8222
Meat	-0.03	0.07	(-0.17, 0.10)	0.8171

Note: 1. The cutting point of the risk indicator is -1.64 FEV1 Z-score.
2. The results are in log scale

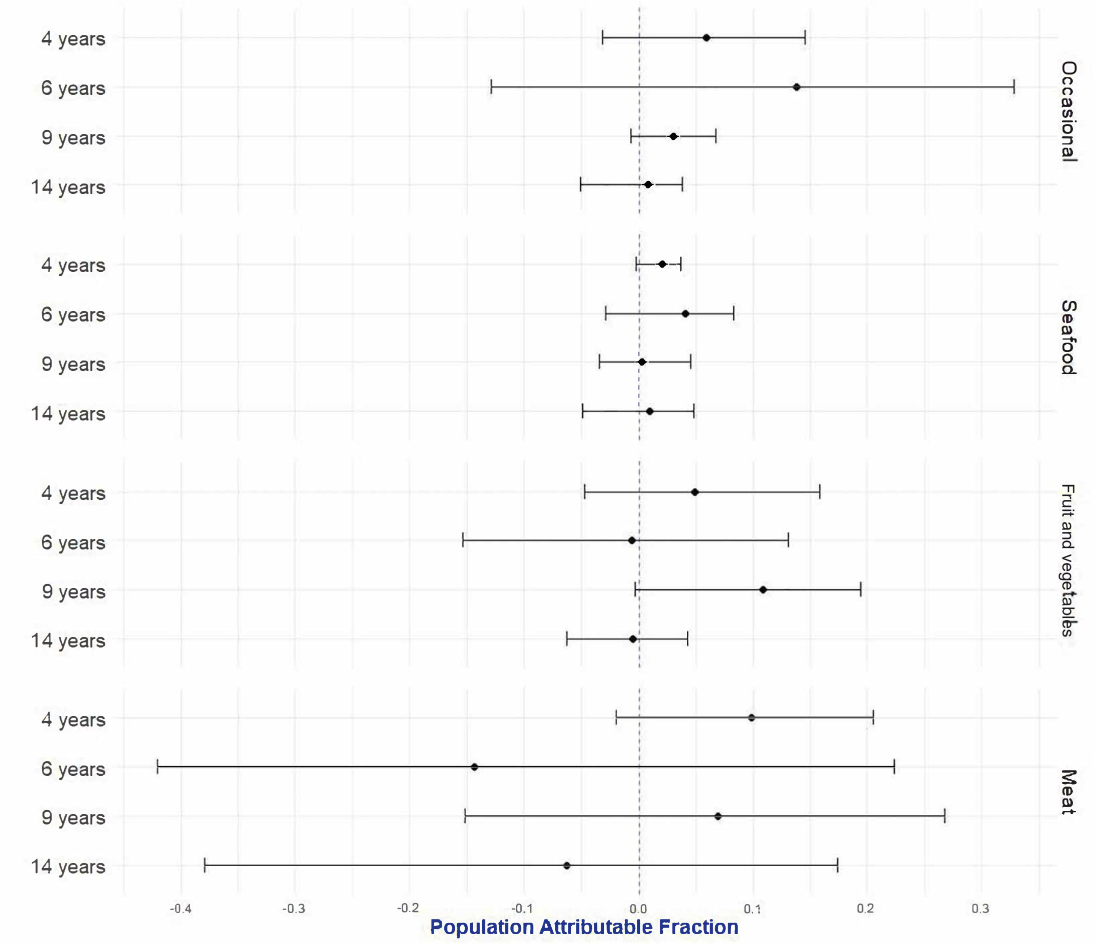
3.6.2 Population attributable fraction

The population attributable fraction (PAF) of nutrition factor scores on the healthy lung function are shown in Table 3.24 and Figure 3.10. Based on the table and the forest plot, the most obvious feature is that none of them is statistically significant as all of their 95% confidence interval contains 0. However, for some eating patterns, we notice that the position of 0 is near to the lower bound in the 95% confidence interval of the estimates, as in the cases of the ”Seafood” eating pattern and the ”Meat” eating pattern at 4 years, and the ”Occasional” and the ”Fruit and vegetables” eating patterns at 9 years showing that they approach significance at the 5% level. Of these, compared to other estimates, the estimated PAF of healthy lung function for the ”Fruit and vegetables” eating pattern at 9 years is remarkably larger than 0. In the light of results from the linear regression model, it may be that a true populational effect is occurring. We estimate that, on average, the current consumption of ”Fruit and vegetables” at 9 years of age has increased healthy lung function prevalence by 11 percentage points, compared to no consumption of ”Fruit and vegetables” at all. The other point to notice is that the forest plot of the PAF of healthy lung function (Figure 3.10) is not consistent with the forest plot of the causal effect of the FEV1 adjusted for height and sex (Figure 3.9). This inconsistency may be caused by the cutting point used to define the healthy lung function, which lies close to an extremity of the range of lung function values.

Table 3.24: Population attributable fraction for nutrition factor scores

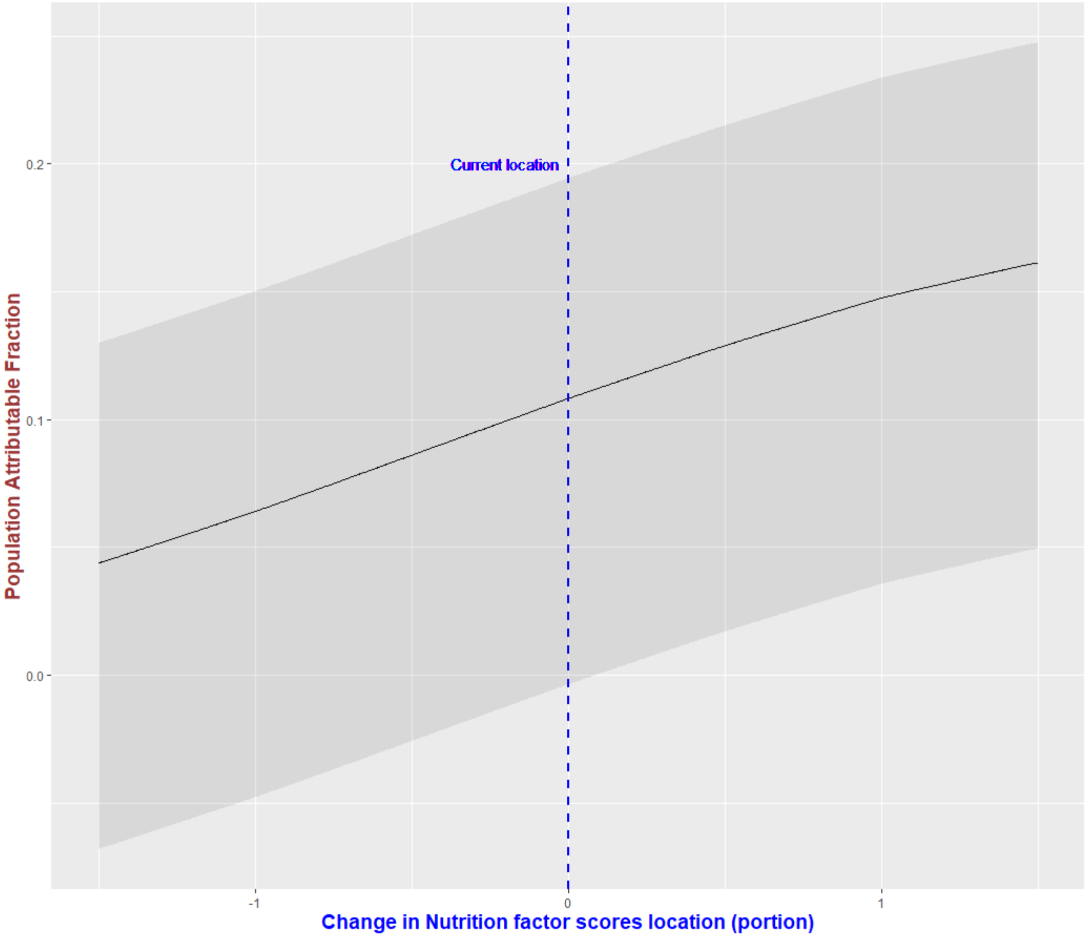
Eating pattern	**Estimate**	**Standard Error**	**95% Confidence Interval**
4 years			
Occasional	0.06	0.05	(-0.03, 0.15)
Seafood	0.02	0.01	(0.00, 0.04)
Fruit and vegetables	0.05	0.05	(-0.05, 0.16)
Meat	0.10	0.06	(-0.02, 0.21)
6 years			
Occasional	0.14	0.12	(-0.13, 0.33)
Seafood	0.04	0.03	(-0.03, 0.08)
Fruit and vegetables	-0.01	0.07	(-0.15, 0.13)
Meat	-0.14	0.16	(-0.42, 0.22)
9-years			
Occasional	0.03	0.02	(-0.01, 0.07)
Seafood	0.00	0.02	(-0.03, 0.05)
Fruit and vegetables	0.11	0.05	(0.00, 0.19)
Meat	0.07	0.11	(-0.15, 0.27)
14 years			
Occasional	0.01	0.02	(-0.05, 0.04)
Seafood	0.01	0.02	(-0.05, 0.05)
Fruit and vegetables	-0.01	0.03	(-0.06, 0.04)
Meat	-0.06	0.13	(-0.38, 0.17)

Figure 3.10: Population Attributable Fraction of healthy lung function



Additionally, we also tried to simulate how the PAF of health lung function changes with nutrition factor score location. Figure 3.11 shows that the PAF of healthy lung function increases monotonically with nutrition factor scores location. This increase is consistent with the possibility of increasing the prevalence of healthy lung function in Pacific youth by increasing the daily intake of ”Fruit and vegetables” at the population level at 9 years of age.

Figure 3.11: The change of PAF of healthy lung function with the different locations of nutrition factor scores in the ”Fruit and vegetables” eating pattern at 9 years



The details of the modified location model is in section 2.6

We note that the location-shift scenario explored here is only an illustration of how changes if distribution in the nutritional exposures can be translated into a change in the prevalence of healthy lung function.

Chapter 4

Discussion

4.1 Summary of findings

In this paper, we aimed to estimate how modifiable risk and protective factors in early life affect the peak lung function and respiratory outcomes in Pacific youth in the early adulthood. Our research is based on the data collected from Pacific Islands Families (PIF) cohort study. According to experts' suggestions, we selected the most relevant risk and protective factors for respiratory health from over 10,000 variables. We then divided them into 11 different domains - Immunisation, Exercise, Breastfeeding, Antenatal smoking, Smoking exposure, Smoking, Respiratory illness-infection, Anthropometrics, Nutrition, Allergies, and Dwelling. However, there were still more than 1300 variables remaining. Therefore, we employed three different methods (Combination of related variables as a single variable, Selection of variables by experts, and Combination of variables by factor analysis) to further reduce the dimensions. Among them, factor analysis is the most complex method and was only utilized for exposures in the nutrition domain. However, before implementing factor analysis, we needed to make food consumption data in the nutrition domain comparable across measurement waves. As the dietary assessment methods are not identical at all measurement waves, the consumption for the same or similar food were measured in different units or with different granularity in different measurement waves. This meant that the food consumption was not comparable amongst measurement waves. We used two steps to resolve this issue : 1. recalculate all food consumption to daily portion; 2. build a mapping to group the same or similar food items or food groups to 12 food categories common to all measurement waves. These 12 food categories were basically consistent with the earlier research [28] but some adjustments were needed.

In the next stage, we reran the Exploratory Factor Analysis (EFA) on all measurement waves to validate whether they had a similar underlying structure (the eating patterns) as the one obtained from the measurement wave at 14 years in the original paper. The results from EFA show that the selected factors and weights have some measure of plausibility across all ages, but align more closely with the eating patterns at 9 and 14 years than at 4 and 6. We persisted with the eating patterns from the early paper as they were strongly supported by nutrition science. Confirmatory factor analysis (CFA) was then introduced to generate the factor loadings for the nutrition factor score computation. To enable the nutrition factor scores to apply on the new data, they had to meet three requirements: 1. nutrition factor scores needed to be invariant over measurement waves; 2. nutrition factor scores needed to be uniformly computed across any data sets; 3. nutrition factor scores needed to be interpretable and open to modification through public health intervention. In this thesis, the first point was achieved by using the measurement invariance model with multi-group CFA, while weighted sum scores, using weights proportional to the loadings, were implemented to satisfy the second and third point, the latter by ensuring that nutrition factor scores were measured in portions per day.

In the third stage, we estimated the causal effect of nutrition factor scores on the respiratory outcomes (FEV1 adjusted for height and sex, FEV1 Z-score, and FEV1 % predicted). We drew the causal diagram to explore the causal paths and identify confounders on the paths. To alleviate the consequences of parametric assumptions (e.g. of normality) semi-parametric linear regression was used. This technique uses a non-parametric approach (boostrap method) to calculate the standard errors. In this way, we could obtain consistent estimates for the causal effects and model-free standard errors, without assuming a true error distribution. The other issue that we needed to resolve before building the model was selection bias in the cohort data. In this thesis, we chose Inverse Probability Weight (IPW) to solve this issue. We applied a logistic model to the indicator of participation into the 18-year wave and of non-missingness of exposure and confounder variables, which was regress on selected baseline characteristics from the birth cohort to generate the weights. The results of the weighting models were reasonable as participants with low probability of participation at 18-year and data completeness were indeed upweighted in the models.

The first models which we fitted were linear regression models. Such models can help us interpret how the respiratory outcomes are, on average, affected by nutrition factor scores (the daily portion of a specific eating pattern). The results of these models reflect that only the ”Fruit and vegetables” eating pattern at 9 years is statistically significant (at the 5% level) and clinically significant at the same time. Based on the linear models, we estimated that, on average, one added portion per day of ”Fruit and vegetables” at 9 years will increase FEV1 Z-score by 0.25 units or FEV1 % predicted by 2.94 percentage points or lung volume by 120 millilitre. We also fitted relative risk models. We used the cutting point (-1.64) of unhealthy lung function in FEV1 Z-score to setup the indicator of healthy lung function as the response for these models. Since we setup the group with participants whose FEV1 Z-score is greater than or equal to -1.64 as the risk group, the results of relative risk model are the ratio of risks based on the beneficial outcome. The relative risk models are transition models used to assist with the calculation of Population attributable fraction (PAF) of protection from the observed distribution of nutrition factor scores at each wave. PAF is a statistic to estimate, at the population level, the proportion of people with the healthy lung function that can be attributed to nutrition factor scores. The results of the PAF analysis show that only the ”Fruit and vegetables” eating pattern at 9 years is statistically significant, which is consistent with the results of linear regression models. This result indicates that the consumption pattern of ”Fruit and vegetables” at 9 years is accountable for 11 percentage points of the prevalence of healthy lung function, compared to o consumption at all. Finally, we showed that the prevalence of healthy lung function can be increased by changing the distribution of the ”Fruit and vegetables” eating pattern at 9 years. This indicates that we can estimate the improvement in healthy lung function prevalence amongst Pacific Island youth by a public health intervention that increases the average consumption of ”Fruit and vegetables” in some predictable way at 9 years.

4.2 Strengths and limitations

Before discussing the strengths of this study, we first talk about its potential limitations:

- Nutrition factor scores used in the study may not be completely compatible with eating patterns from all measurement waves. This is because we tried to keep the eating patterns used in this study aligned with eating patterns identified in early research, but the findings were only based on data from the 14-years measurement wave. Therefore, there may be compatibility issues with other measurement waves. Lack of association between nutrition factor scores and respiratory outcomes in the early measurement waves may be due in part to this incompatibility.

- The questionnaires were not uniform across all measurement waves in the PIFS cohort. The latter two measurement waves utilized different questionnaires from the early two measurement waves. To unify the food categories amongst all measurement waves, we dropped some food items only asked in the early two measurement waves. This may lead to some information being lost in the study.

- There may be some missing exposures in the causal diagram. This means that the causal diagram may not be able to present the complete picture of causal paths to healthy lung function, which may lead to some bias in the results of the causal model.

Some strengths also should be highlighted in our study:

- In this study, we unified the unit of nutrition factor scores over all measurement waves to make them comparable.

- As nutrition factor scores were expressed in daily portion, the PAF obtained in our study actually revealed how the prevalence of healthy lung function can be affected by a change in the number of daily portions of food consumption in the particular eating pattern. It offers a feasible solution for public health intervention to enhance the lung function of Pacific Island youth.

- The causal diagram was reviewed by experts in the relevant areas, and every domain identified as important in terms of exposure or confounding is represented. Therefore, we are confident that the causal models fitted in the study shall cover most of the causal paths between nutrition factor scores and the respiratory outcomes, meaning results should be nearly unbiased.

- We applied inverse probability weight (IPW) on all the models in the study. The weighting generated by IPW may not be the best, but it can to a certain degree compensate for the impact of the selection bias.

4.3 Future research

In future research, we can improve our findings in the following aspects: 1. rerun the factor analysis on the 9-years measurement wave without following the early paper, and rebuild the models and recompute the PAF based on new nutrition factor scores to see whether the results agree with our findings; 2. use other methods to calculate the weights, such as implementing multiple imputation to impute missing values of exposures and confunders and then calculate uniform wights for all eating patterns in a measurement wave. Another refinement would consist in modelling the probability of participation at each measurement wave using all data available up that wave, and obtain a final probability of participation at 18-years through the product of these probabilities; 3. examine various ways to change the distribution of the location of nutrition factor scores to reveal how the PAF of healthy lung function varies over different situations.

References

[1] L. Riley and M. Cowan, “World health organization noncommunicable diseases country profiles,” *Geneva, Switzerland: WHO Library Cataloguing-in-Publication Data*, 2014.

[2] S. N. Zealand, “2018 census: Stats nz.” `https://www.stats.govt.nz/2018-census/`, 2021.

[3] S. N. Zealand, “National ethnic population projections: 2018 (base)–2043,” *Wellington Statistics New Zealand*, 2021.

[4] Asthma and R. F. of New Zealand, “Te hā ora (the breath of life): National respiratory strategy,” 2015.

[5] L. Telfar Barnard, M. Baker, N. Pierse, and J. Zhang, *The impact of respiratory disease in New Zealand: 2014 update*. in April 2015 by the Asthma Foundation., 2015.

[6] W. D. Lees J, Lee M, *Demographic Profile: 2018 Census Population of Counties Manukau*. Counties Manukau Health, 2021.

[7] J. Paterson, T. Percival, P. Schluter, G. Sundborn, M. Abbott, S. Carter, E. Cowley-Malcolm, J. Borrows, W. Gao, and P. S. Group, “Cohort profile: the pacific islands families (pif) study,” *International Journal of Epidemiology*, vol. 37, no. 2, pp. 273–279, 2008.

[8] G. Sundborn, J. Paterson, U. Jhagroo, S. Taylor, L. Iusitini, E.-S. Tautolo, A. H. Fa’asisila Savila, and M. Oliver, “Cohort profile: A decade on and strong-the pacific islands families study,” *AUT Pacific Islands Families Study Of those Born in 2000, at Manukau City, New Zealand*, vol. 17, no. 2, p. 9, 2011.

[9] E. Rush, M. Oliver, L. Plank, S. Taylor, L. Iusitini, S. Jalili-Moghaddam, F. Savila, J. Paterson, and E. Tautolo, “Cohort profile: pacific islands families (pif) growth study, auckland, new zealand,” *BMJ open*, vol. 6, no. 11, p. e013407, 2016.

[10] J. Stocks, A. Hislop, and S. Sonnappa, “Early lung development: lifelong effect on respiratory health and disease,” *The lancet Respiratory medicine*, vol. 1, no. 9, pp. 728–742, 2013.

[11] A. Bush, “Copd: a pediatric disease,” *COPD: Journal of chronic obstructive pulmonary disease*, vol. 5, no. 1, pp. 53–67, 2008.

[12] O. Savran and C. S. Ulrik, “Early life insults as determinants of chronic obstructive pulmonary disease in adult life,” *International journal of chronic obstructive pulmonary disease*, vol. 13, p. 683, 2018.

[13] E.-S. Tautolo, C. Wong, A. Vandal, S. Jalili-Moghaddam, E. Griffiths, L. Iusitini, A. Trenholme, and C. Byrnes, “Respiratory health of pacific youth: An observational study of associated risk and protective factors throughout childhood,” *JMIR Research Protocols*, vol. 9, no. 10, p. e18916, 2020.

[14] N. Soto-Ramírez, M. Alexander, W. Karmaus, M. Yousefi, H. Zhang, R. J. Kurukulaaratchy, A. Raza, F. Mitchell, S. Ewart, and S. H. Arshad, “Breastfeeding is associated with increased lung function at 18 years of age: a cohort study,” *European respiratory journal*, vol. 39, no. 4, pp. 985–991, 2012.

[15] S. Filoche, S. Garrett, J. Stanley, S. Rose, B. Robson, C. R. Elley, and B. Lawton, “Wāhine hauora: linking local hospital and national health information datasets to explore maternal risk factors and obstetric outcomes of new zealand māori and non-māori women in relation to infant respiratory admissions and timely immunisations,” *BMC Pregnancy and Childbirth*, vol. 13, no. 1, pp. 1–6, 2013.

[16] M. J. Herring, L. F. Putney, G. Wyatt, W. E. Finkbeiner, and D. M. Hyde, “Growth of alveoli during postnatal development in humans based on stereological estimation,” *American Journal of Physiology-Lung Cellular and Molecular Physiology*, vol. 307, no. 4, pp. L338–L344, 2014.

[17] J. P. Butler, S. H. Loring, S. Patz, A. Tsuda, D. A. Yablonskiy, and S. J. Mentzer, “Evidence for adult lung growth in humans,” *New England Journal of Medicine*, vol. 367, no. 3, pp. 244–247, 2012.

[18] M. Narayanan, J. Owers-Bradley, C. S. Beardsmore, M. Mada, I. Ball, R. Garipov, K. S. Panesar, C. E. Kuehni, B. D. Spycher, S. E. Williams, *et al.*, “Alveolarization continues during childhood and adolescence: new evidence from helium-3 magnetic resonance,” *American journal of respiratory and critical care medicine*, vol. 185, no. 2, pp. 186–191, 2012.

[19] E. T. Zemanick, J. K. Harris, S. Conway, M. W. Konstan, B. Marshall, A. L. Quittner, G. Retsch-Bogart, L. Saiman, and F. J. Accurso, "Measuring and improving respiratory outcomes in cystic fibrosis lung disease: opportunities and challenges to therapy," *Journal of Cystic Fibrosis*, vol. 9, no. 1, pp. 1–16, 2010.

[20] D. R. VanDevanter and M. W. Konstan, "Outcome measures for clinical trials assessing treatment of cystic fibrosis lung disease," *Clinical investigation*, vol. 2, no. 2, p. 163, 2012.

[21] D. E. Niewoehner, D. Collins, and M. L. E. f. t. D. o. Veterans Affairs Cooperative Study Group, "Relation of fev1 to clinical outcomes during exacerbations of chronic obstructive pulmonary disease," *American journal of respiratory and critical care medicine*, vol. 161, no. 4, pp. 1201–1205, 2000.

[22] M. Cazzola, W. MacNee, F. Martinez, K. F. Rabe, L. Franciosi, P. Barnes, V. Brusasco, P. Burge, P. Calverley, B. Celli, *et al.*, "Outcomes for copd pharmacological trials: from lung function to biomarkers," *European Respiratory Journal*, vol. 31, no. 2, pp. 416–469, 2008.

[23] A. Augarten, H. Akons, M. Aviram, L. Bentur, H. Blau, E. Picard, J. Rivlin, M. S. Miller, D. Katznelson, A. Szeinberg, *et al.*, "Prediction of mortality and timing of referral for lung transplantation in cystic fibrosis patients," *Pediatric transplantation*, vol. 5, no. 5, pp. 339–342, 2001.

[24] W. Robinson and D. A. Waltz, "Fev1 as a guide to lung transplant referral in young patients with cystic fibrosis," *Pediatric pulmonology*, vol. 30, no. 3, pp. 198–202, 2000.

[25] J. Courtney, J. Bradley, J. Mccaughan, T. O'connor, C. Shortt, C. Bredin, I. Bradbury, and J. Elborn, "Predictors of mortality in adults with cystic fibrosis," *Pediatric pulmonology*, vol. 42, no. 6, pp. 525–532, 2007.

[26] R. Young, R. Hopkins, and T. Eaton, "Forced expiratory volume in one second: not just a lung function test but a marker of premature death from all causes," *European Respiratory Journal*, vol. 30, no. 4, pp. 616–622, 2007.

[27] J. F. Donohue, "Minimal clinically important differences in copd lung function," *COPD: Journal of Chronic Obstructive Pulmonary Disease*, vol. 2, no. 1, pp. 111–124, 2005.

[28] S. Jalili Moghaddam, *What Children are Eating and the Risk of Type 2 Diabetes Mellitus*. PhD thesis, Auckland University of Technology, 2018.

[29] K.-C. Li, "Sliced inverse regression for dimension reduction," *Journal of the American Statistical Association*, vol. 86, no. 414, pp. 316–327, 1991.

[30] R. D. Cook and S. Weisberg, *An introduction to regression graphics*, vol. 405. John Wiley & Sons, 2009.

[31] C. DiStefano, M. Zhu, and D. Mindrila, "Understanding and using factor scores: Considerations for the applied researcher," *Practical Assessment, Research, and Evaluation*, vol. 14, no. 1, p. 20, 2009.

[32] T. A. Brown, *Confirmatory factor analysis for applied research*. Guilford publications, 2015.

[33] W. Revelle and M. W. Revelle, "Package 'psych'," *The comprehensive R archive network*, vol. 337, p. 338, 2015.

[34] J. Reinecke and A. Pöge, *Confirmatory Factor Analysis*. SAGE Publications Limited, 2020.

[35] H. Steinmetz, P. Schmidt, A. Tina-Booh, S. Wieczorek, and S. H. Schwartz, "Testing measurement invariance using multigroup cfa: Differences between educational groups in human values measurement," *Quality & Quantity*, vol. 43, no. 4, pp. 599–616, 2009.

[36] Y. Rosseel, "Lavaan: An r package for structural equation modeling and more. version 0.5–12 (beta)," *Journal of statistical software*, vol. 48, no. 2, pp. 1–36, 2012.

[37] J. Tian and J. Pearl, "A general identification condition for causal effects," in *Aaai/iaai*, pp. 567–573, 2002.

[38] J. Textor, B. van der Zander, and M. J. Textor, "Package 'dagitty'," 2016.

[39] M. Barrett, "ggdag: Analyze and create elegant directed acyclic graphs," *R package version 0.1. 0*, 2018.

[40] S. Stanojevic, A. Wade, and J. Stocks, "Reference values for lung function: past, present and future," *European Respiratory Journal*, vol. 36, no. 1, pp. 12–19, 2010.

[41] C. Wong. Private communication, 2021.

[42] M. A. Hernán, S. Hernández-Díaz, and J. M. Robins, "A structural approach to selection bias," *Epidemiology*, pp. 615–625, 2004.

[43] J. Ferguson, F. Maturo, S. Yusuf, and M. O'Donnell, "Population attributable fractions for continuously distributed exposures," *Epidemiologic Methods*, vol. 9, no. 1, 2020.

[44] C. J. Lloyd, "Estimating attributable response as a function of a continuous risk factor," *Biometrika*, vol. 83, no. 3, pp. 563–573, 1996.

Appendix A

Supplementary Figure

A.1 The distribution of inverse probability weights (IPW)

Figure A.1: The distribution of inverse probability weights (IPW) for the Occasional eating pattern across measurement waves

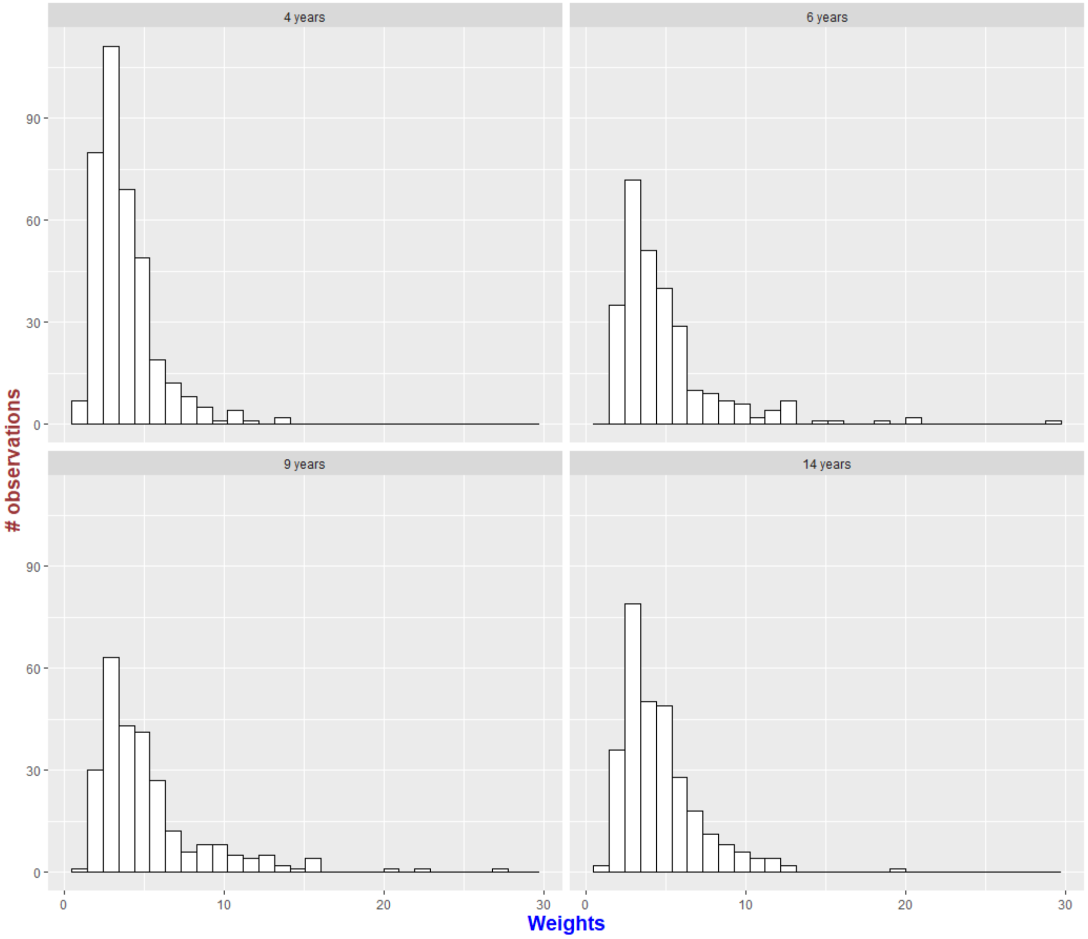


Figure A.2: The distribution of inverse probability weights (IPW) for the Seafood eating pattern across measurement waves

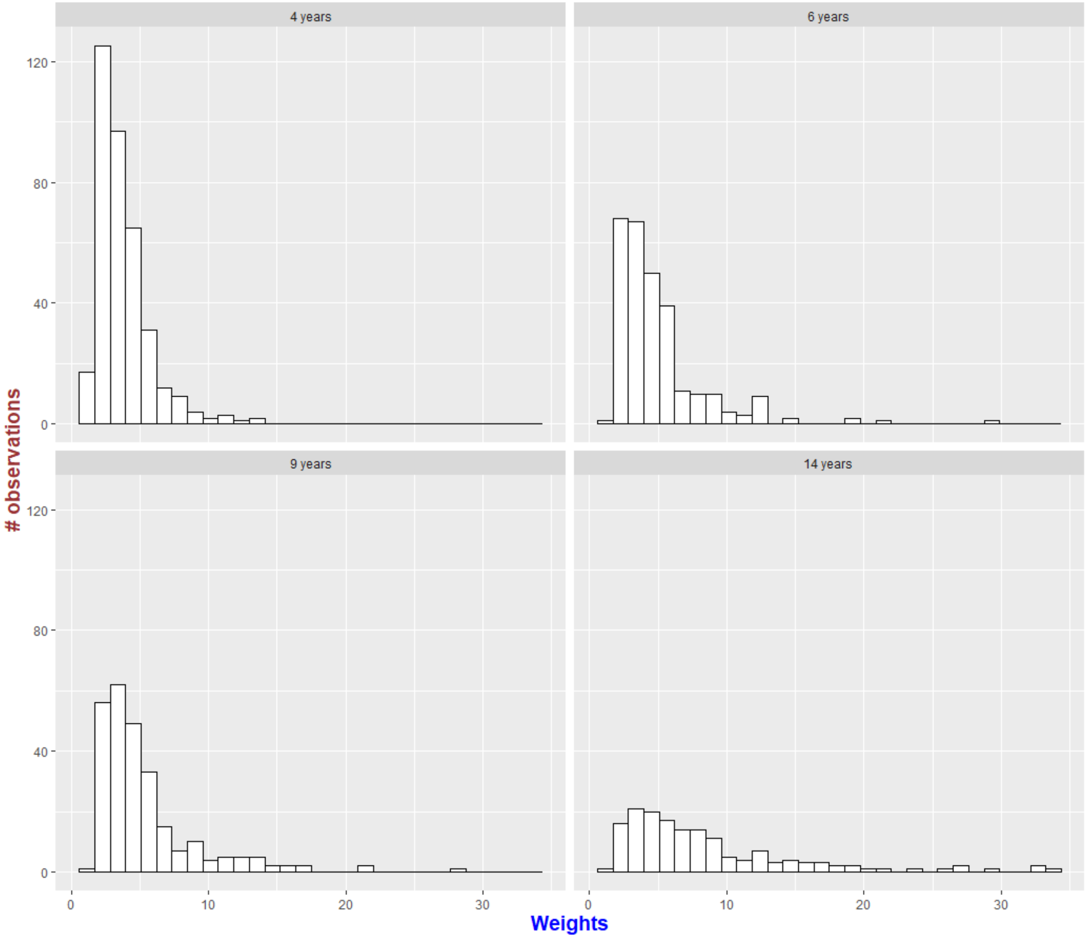
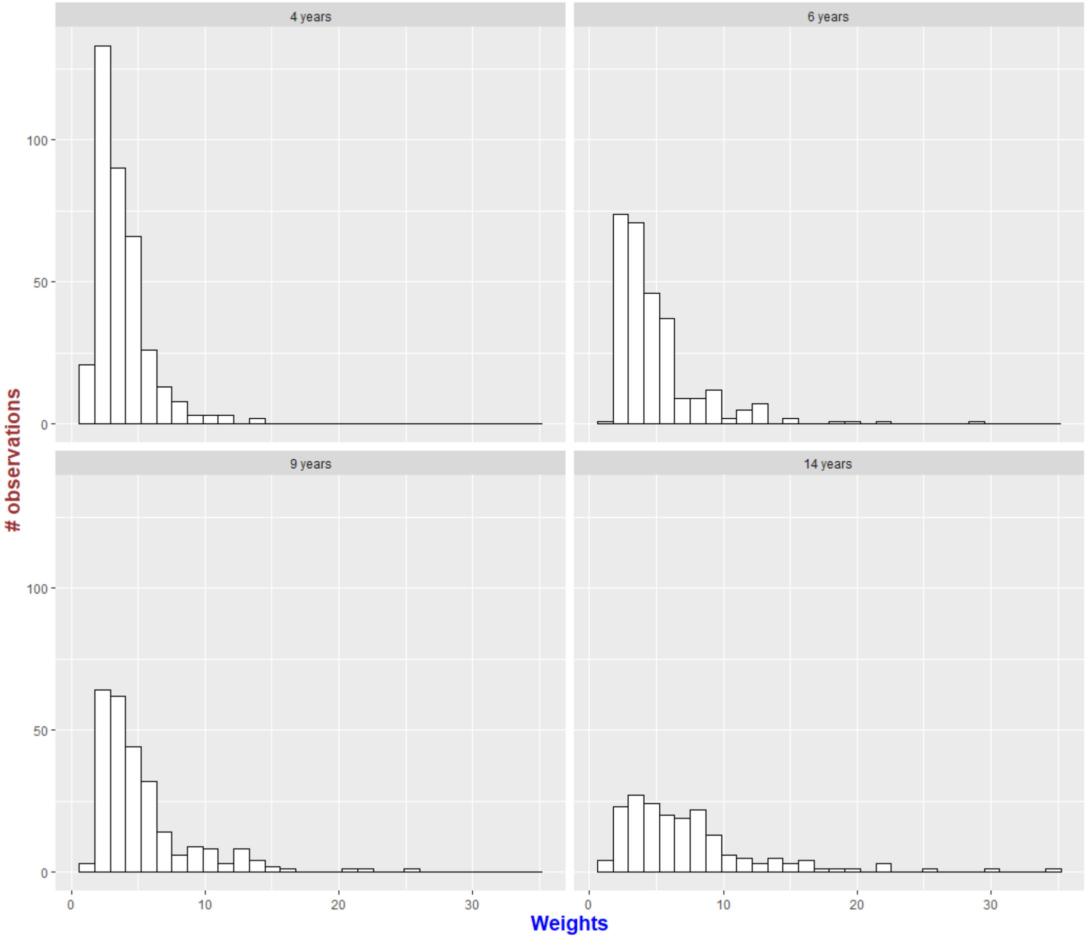


Figure A.3: The distribution of inverse probability weights (IPW) for the Meat eating pattern across measurement waves


4 years
6 years
9 years
14 years
# observations
Weights