Epidemiological evidence that can help to improve timely diagnosis of colorectal cancer in New Zealand

A thesis

submitted in fulfilment

of the requirements for the degree

of

Doctor of Philosophy in Health Development and Policy

at

The University of Waikato

by

Malgorzata Hirsz

THE UNIVERSITY OF
WAIKATO
Te Whare Wānanga o Waikato

2021

Abstract

Incidence rates of colorectal cancer (CRC) in New Zealand (NZ) are among the highest in the world. The long-term survival rates in NZ are poor, which has partially been attributed to late diagnosis. Considering that CRC is a curable disease if diagnosed early, conducting research targeted at the improvement of early diagnosis is therefore important. This thesis provides statistical models for the calculation of CRC incidence rates in the entire NZ population and population strata defined by age, gender, ethnicity and diabetes status, and a model for CRC risk in individual patients referred to the secondary care. The models presented here could assist health professionals in the selection of patients for further investigation to facilitate earlier diagnosis. The empirical part consists of three independent observational studies, briefly described below.

Sub-study 1. The objectives were, first, to describe trends in CRC incidence in the NZ population and, second, to investigate whether there are any strata defined by gender and ethnicity with especially increased incidence rates of CRC. To address these objectives, I analysed data from the New Zealand Cancer Registry (years 1994–2018) using an age-period-cohort (APC) model. The overall CRC incidence rates in NZ decreased between 1994 and 2018 by an average of 1.31% per year. However, the decrease was observed only in patients 50 years and older. In those 30-<50 years old, the incidence rates increased between 1994 and 2018 regardless of gender and ethnicity. The increase was similar for proximal, distal and rectal cancers. The APC analyses revealed very strong cohort effects that could explain nearly the entire trends in CRC incidence, pointing out generations born in the 1970s and 80s being affected by the increased incidence rates, rather than individuals 30-<50 years old. The cohort effects were different in Māori and non-Māori populations. In non-Māori born between approximately 1939 and 1955, incidence rates decreased sharply. By contrast, those Māori generations have not benefitted from the sharp decrease in rates. However, CRC incidence increased substantially in both Māori

and non-Māori groups born in the 1970s and 80s.

Sub-study 2. The objective was to estimate the IRR for CRC in patients with diabetes compared to those without diabetes, with relation to diabetes duration and use of insulin for diabetes control. Registration in the Virtual Diabetes Register (VDR) in the years 2014–2018 was used as a marker of a diabetes diagnosis. Tables with counts of the entire NZ population stratified by age, gender, and ethnicity were obtained from Statistics NZ. In total, data from 310,710 patients with diabetes, corresponding to 1,277,284 person-years and 2512 incident CRC cases were analysed using a Poisson regression model. Diabetes was associated with an overall increased CRC incidence of 13% compared to non-diabetes [IRR=1.13 (95% CI: 1.08, 1.18)]. The IRR was especially high in the first three months after diabetes diagnosis [IRR=2.55 (95% CI 2.02, 3.21)], likely due to detection bias. The association was equally strong in males and females. However, in the analysis by ethnicity, the incidence of CRC was increased only in non-Māori patients and restricted only to those younger than 75 years.

Sub-study 3. The objective was to develop a predictive model for CRC risk in individual patients referred to secondary care. To develop such a model, I extracted information from free text included in e-referrals from GPs' to the Gastroenterology and General Surgery departments in the Waikato Hospital from 2015-2018, including: symptoms; test results; and family history of CRC. The reference test was a full colonoscopy with visualisation of the cecum. Data from 3015 patients, 20-<90 years old were analysed using a logistic regression model. The final model included the following predictors associated with increased CRC risk: anaemia, rectal bleeding, palpable mass in abdomen or rectum, weight loss, age and gender, and a decreased CRC risk: family history of CRC, abdominal pain, and inflammatory bowel disease. The model discriminates patients with low CRC risk well. According to the final cross-validated model, around 20% of patients from our cohort had performed colonoscopy despite a very low CRC risk (less than 1.5%).

In conclusion, the APC analysis revealed an alarming pattern. According to the fitted APC model, the combination of increasing age and cohort effect in generations born in the 1970s and 80s will bring a wave of CRC diagnoses in the near future when the young generations with high CRC incidence rates will replace the old generations with low CRC incidence. The results from this study could therefore help policy-makers to plan the needs for gastroenterology services.

Secondly, CRC incidence rates in diabetes have been found to be slightly increased compared to non-diabetes but only in non-Māori individuals younger than 75 years.

Third, the results suggest that males underutilise health services. As shown in sub-study 3, males underwent fewer colonoscopies than females, despite having a higher risk of CRC. The higher detection bias in males than in females (sub-study 2) could also suggest underuse of health services by males, but the difference was not statistically significant.

Finally, based on the fitted models for CRC incidence in sub-studies 1 and 2, population-wide CRC screening for Māori and patients with diabetes, based on the incidence, instead of age alone, would be proposed to start at age 57.5 years if the screening in the general population starts at age 60 years.

Preface

This PhD thesis combines my love for health research with the knowledge about research methods used in epidemiology which I gained over the three-year PhD study building on my background education including: a BSc in Public Health Nutrition, and MSc in Applied Statistics. Here, I present the work which I hope can contribute to the improvement of early diagnosis of colorectal cancer patients in New Zealand, a beautiful country that has been my home for the last four years.

Malgorzata Hirsz
29 March 2021

Acknowledgements

Carrying out the study, and writing this thesis, has been a journey, and there are several people whom I would like to thank for their help along the way.

Firstly, I would like to thank my supervisors, Lyn Hunt, Lynne Chepulis, and Michael Mayo, for their guidance, support, for always being there for me when I needed, and for the freedom they gave me to explore gaps in the existing knowledge about the epidemiology of colorectal cancer in New Zealand. I have learnt a lot from their valuable feedback.

Thanks must also go to Professor Ross Lawrenson, who introduced me to the research on colorectal cancer in New Zealand, supported my application for the University of Waikato Doctoral Scholarship and provided mentorship at the start of this PhD. Professor Lawrenson also granted me access to the electronic referrals data, collected for the HRC study purposes.

I would like to acknowledge Helene Høgsbro Thygesen, for ongoing discussions about statistical methodology and difficult concepts in epidemiology, for help in writing some of the more complex R-codes, and for tips on the management of large data sets. The multiple rounds of critical but supportive feedback she provided shaped my understanding of many problems which arrived during the study.

All this shaped my approach to health research in deep ways and has helped me to develop my own ideas and methods throughout this PhD project.

I am grateful to all academic staff of the Waikato University who made the 3-year journey such a beautiful experience (even in the Covid-19 era), with special thanks going to my two tutors: Nicole Pepperell for teaching the importance of ethics in academia and for sharing her wisdom; and Chelsea Blickem for inspirational teaching of skills to present

research to a wide audience. Thanks also to all my colleagues for their great company, especially to Nick Lim for help with solving issues related to technical problems with using LaTeX.

I am grateful to the University of Waikato for providing financial support through the University of Waikato Doctoral Scholarship award, and for providing support with research expenses that allowed me to attend scientific conferences and to pay for the acquisition of data from the Ministry of Health.

My thanks go also to the data managers from the Ministry of Health for their help with decisions about the type and cost of specific data, and the excellent preparation of the agreed data sets, as well as to Statistics NZ for help with access to population counts for the research purposes.

Last but not least, thanks to Duncan Law for proofreading the entire thesis.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

AIC	Akaike's Information Criterion
AAPC	Average annual percentage change
APC	Age-period-cohort
ASR	Age-standardised rate
AUC	Area under the ROC curve
BIC	Bayesian Information Criterion
CRC	Colorectal cancer
DM	Diabetes mellitus
FOBT	Faecal occult blood test
FSA	First specialist assessment
GP	General practitioner
HR	Hazard ratio

HRC	Health Research Council
IR	Incidence rate
IRR	Incidence rate ratio
MoH	Ministry of Health
NBSP	National Bowel Screening Programme
NHI	National Health Index
NZCR	New Zealand Cancer Registry
OR	Odds ratio
ROC	Receiver operating characteristic
VDR	Virtual Diabetes Registry
WHO	World Health Organisation

List of statistical terms

Crude incidence rate

The total number of cases divided by the corresponding population count expressed as cases per 100,000 person years.

Age-specific incidence rate

Incidence rates (crude or model-based) calculated/estimated for each age-group.

Age-standardised incidence rate

The incidence rate standardised according to the age distribution of a given reference population. In this study the structure of the New Zealand population 2018 was used. The method for calculating the age standardised incidence rate is described in Boniol and Heanue (2007).

Cumulative probability of developing cancer

The probability of developing cancer in a certain age range Boniol and Heanue (2007) (in this study between 30-<90 years assuming no competing risks).

Trend in incidence

Is provided as the average annual percentage change based on net drift extracted from age-period-cohort model.

Chapter 1

Introduction

1.1 Introduction

This thesis is concerned with providing epidemiological evidence related to colorectal cancer (CRC) risk and CRC incidence rates in the New Zealand (NZ) population, based on analyses of existing population-based and administrative data sets. NZ is facing a problem with one of the highest incidence rates of CRC in the world, accompanied by low long-term survival rates by international standards. The poor survival is partially caused by the diagnosis of a high proportion of patients in advanced stages when long-term survival is unlikely. Considering that CRC is one of the most curable cancers if diagnosed early, improvement of timely diagnosis is a promising strategy to improve the long term survival and therefore research on strategies for early diagnosis in needed. In NZ, more patients than in other developed countries are diagnosed with CRC after the presentation to Emergency Departments (ED) which suggests that some patients may be missed in the diagnostic pathway by primary or secondary care doctors, or some patients might not recognise the seriousness of the symptoms and therefore not report them to GPs. The delay in diagnosis of CRC can therefore be related to patient or system factors. This thesis does not investigate patient factors and provides only results that can be useful to address delay in diagnosis related to system factors[1].

To chose objectives for this study, I firstly addressed the following question: what type of new evidence-based on NZ data could help in addressing the CRC burden? Based on research conducted in other countries, I found that analysis of data from cancer registries and other population-based sources can reveal useful information for dealing with the CRC burden. Importantly, NZ has an excellent cancer registry and population-based data sources that can be used to generate valuable evidence, but the data have until now been under-utilised for CRC epidemiology research.

[1]The results provided by this study can assist in addressing the delay in diagnosis that occurs in secondary care or the delay occurring due to the policy for screening eligibility, and possibly in primary care.

To illustrate the motivation of this study, in Figure 1.1, I introduce a fictional patient called Mary in a typical situation within one of the steps of the diagnostic pathway, here in the primary care setting during an appointment with her GP.

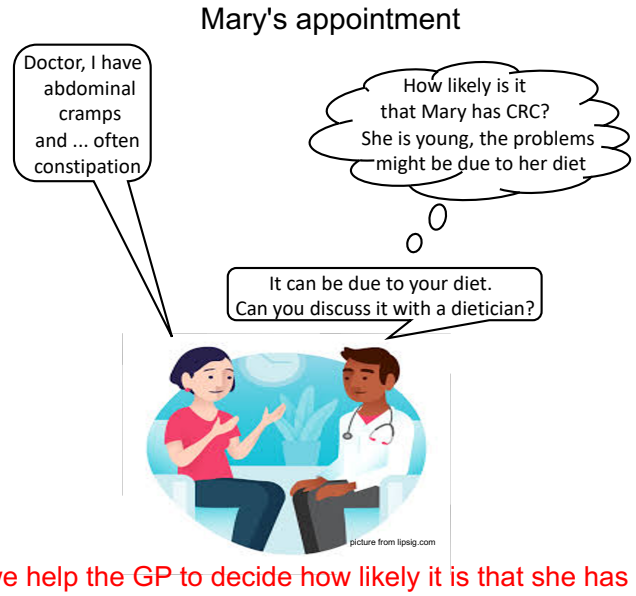


Figure 1.1: *A fictional patient as an example of an individual who will need assessment of the likelihood of CRC.*

During the appointment, the GP has to assess how likely it is that Mary has CRC. Many patients with lower abdominal symptoms visit GPs, but nearly all of them have only non-malignant diseases. Because a GP sees on average only one patient with CRC per year, it is difficult for the GPs to get experience that allows them to identify, among many patients, those who are sufficiently likely to have CRC. For the purpose of this study, I assumed that any additional information about associations between potential risk factors and CRC[2] can help physicians to assess how likely it is that a given patient has

[2]In this study, in analyses which provided results relevant to the primary care and policy-makers the following risk factors were investigated: age, gender, ethnicity and diabetes status. Models relevant for the use in secondary care included age, gender and symptoms stated in the referrals of patients to the specialists.

CRC and whether a prompt investigation for presence/absence of CRC is needed.

The remainder of this chapter outlines the entire project by presenting: the rationale and the significance of this study; the aim and objectives of this study; the methodological framework; and finally, the thesis structure. Unconventionally, the literature that gives background knowledge about the epidemiology of CRC and the literature that led to the identification of gaps in CRC epidemiology in NZ which this research addressed is reviewed in the following chapter (Chapter 2).

1.2 Rationale and significance

Colorectal cancer is a malignant disease that is a cause of premature deaths and a burden to the whole world population (Favoriti et al., 2016). Despite the fact that CRC is a leading cause of cancer mortality, it is a curable disease if diagnosed early (Cheah, 2009); an important characteristic of CRC that motivates worldwide cancer surveillance research to provide evidence that can help in earlier diagnosis. Cancer surveillance research is a branch of cancer epidemiology which uses population-based data such as cancer registries to provide cancer statistics about incidence trends and cancer outcomes (Glaser et al., 2005). Cancer surveillance research specific to CRC covers a broad spectrum of topics including research on risk factors, symptoms presented in patients with CRC and analysis of incidence rates over long periods. The wide spectrum of possible application of results provided by such studies includes: use in CRC control e.g. to manage the available resources; improvement of the diagnostic pathway; and can also provide etiologic clues about causal factors in cancer development, valuable knowledge for public health education and cancer prevention.

However, in NZ, a country with one of the highest incidences of CRC in the world accompanied with poor long-term survival rates by international standards, cancer surveillance

research concerning CRC incidence rates and factors associated with CRC incidence is scarce. Especially, the knowledge about the trends in CRC incidence by ethnicity is limited despite the fact that the disparities in CRC incidence between Māori and non-Māori are frequently discussed in the NZ literature. Specifically, there is no published research which assessed if the differences between the incidence trends in Māori and non-Māori are related to the effect of year of diagnosis or to the effect of birth cohort. Research from other countries has shown that using age-period-cohort analysis, an advanced statistical tool used in cancer surveillance research to analyse cancer registry data, can answer such questions. NZ has good quality cancer registry data and also the best data related to cancer for the Indigenous population in the world (Gurney et al., 2020) available for research. Analysis of those data gives an excellent opportunity to provide valuable information that can advance our knowledge and, possibly, aid in the improvement of earlier diagnosis of CRC in NZ, thus help in addressing the CRC burden.

This study analysed CRC incidence data from the New Zealand Cancer Registry (NZCR) years 1994–2018, using age-period-cohort models; it described long-term trends in CRC incidence for the whole population, as well as separately for Māori and non-Māori. The description of long-term trends of CRC in chapter 3, with disentangled effects of age at diagnosis, year of birth, and year of diagnosis on incidence rates, for both Māori and non-Māori, provides results that have not been published before. In addition to enriching our knowledge about the epidemiology of CRC in NZ which is also of interest for the international research, this study provides model-based incidence rates for population strata defined by age and gender, separate for anatomical sub-sites (for proximal, distal and rectal cancers), and also model-based rates by age and ethnicity. The predicted incidence rates have potential use in clinical practice: in policy-making for revision of referral guidelines, for decisions about the allocation of resources for surveillance of strata at increased risk, for targeting of awareness campaigns, and decisions about possible stratification for population CRC screening; and in primary care where GPs can use the updated knowl-

edge about CRC incidence rates in population strata during decision making for further investigation of individual patients to confirm or rule out cancer [3].

In addition to individual patients' demographic characteristics, comorbidities may play a significant part in CRC incidence. In this study, I have chosen to investigate the association between diabetes and CRC for the following reasons: firstly, there is a strong evidence for a positive association between those two diseases (Tsilidis et al., 2015) however with differences between countries and between ethnicities (González et al., 2017); secondly, the CRC incidence in NZ is one of the highest in the world, and diabetes is a prevalent condition in the NZ population, especially in Māori, (Sundborn et al., 2007) which makes the investigation of the association between diabetes and CRC important for the whole population but especially for Māori; thirdly, to the best of my knowledge the association has not yet been studied in the NZ population; and finally, the availability of population based data on diabetes status in the NZ population made the investigation feasible during the time available for this study. This study, therefore, contributes to the NZ epidemiology of CRC, and also to the international research in this area. Within the NZ setting, an estimate of the strength of the association between CRC and diabetes can help policy-makers to decide whether population-based screening for CRC should be stratified by diabetes status and to update doctors' knowledge about diabetes as a risk factor in the NZ population so doctors do not rely on estimates based on data from other populations.

In NZ, the final investigation for the presence/absence of CRC in symptomatic patients is carried out mostly by specialists in secondary care settings, usually using a colonoscopy. Due to the limited resources for colonoscopy in NZ, specialists have to select patients referred by GPs for further investigation. Statistical models for selection of patients for colonoscopy, based on CRC risk, can support specialists in the identification of patients

[3]The idea of using results from epidemiological studies for the assessment of individual patients is based on Sackett et al. (1985).

with CRC risk high enough to justify a colonoscopy (Adelstein et al., 2010). However, there is no published model for calculation of CRC risk in the population referred to secondary care, based on data collected from NZ patients. This study, therefore, developed a model for calculation of CRC risk, using data included in referrals to the gastroenterology and general surgery departments in the Waikato Hospital in the years 2015–2018. The proposed model can discriminate patients with low CRC risk, for whom colonoscopy is not necessary, and thereby help to achieve a justified reduction in negative colonoscopies performed on patients with low risk of CRC. Those saved colonoscopies could then, for example, be offered to investigate promptly patients with higher risk. In NZ, the proposed model after external validation using data from other hospitals, could help in better management of the limited colonoscopy resources.

The results of the research presented in this thesis are relevant to medical doctors, policy-makers and epidemiologists in NZ; moreover, many of the results presented here can make a valuable contribution to international CRC research. Within NZ, the evidence could be useful to various stakeholders concerned with policies and guidelines for the improvement of the early diagnosis of CRC in NZ in primary care, in secondary care and at assisting policy-making with respect to stratification of the population for CRC screening, and updating policy related to CRC diagnosis.

1.3 Study aim and objectives

The overall aim of the study was to provide new epidemiological evidence, relevant to the improvement of early diagnosis of CRC in NZ, and narrowed down to the following three objectives:

1. Analysis of trends in CRC incidence in the NZ population and identification of population strata with especially increased incidence rates.

2. Estimation of the association between diabetes mellitus and CRC in the NZ population, with relation to the duration of diabetes and insulin use.

3. Fitting a predictive model for estimating CRC risk in individual patients, based on symptoms indicated in e-referral data and on patients' characteristics, for use by clinicians in secondary care settings for selection of patients for colonoscopy.

1.4 Study methodology

This section outlines the methodology used in this study to address the above objectives. In the sub-sections below, I discuss: the study design; the population studied and the data sources; the outline of the general statistical approach; and the ethical clearance for each part of the study. Two main issues were considered when selecting the relevant methods for addressing the study's objectives: first, the feasibility of carrying out the study during the three-year PhD time-frame and with limited financial resources; and, second, the importance of providing estimates with narrow confidence intervals, given that the results are intended for use in clinical practice. The reporting was based on the STROBE recommendations (Vandenbroucke et al., 2014), also guided by the TRIPOD statement Moons et al. (2015). In the process of development and validation of predictive models, I made extensive use of principles described in (Steyerberg et al., 2019).

1.4.1 Study design and study population

This study has a non-experimental design, analysing already existing data sets; population-based registry data and administrative data. Population-based data sets are valuable sources of information for scientific research, due to minimal selection bias and the applicability of findings to the whole population (Gavrielov-Yusim and Friger, 2014).

These attributes are especially important when the research is intended for use in clinical practice.

The empirical part of this PhD project consisted of three separate sub-studies, which will be referred to as sub-study 1, sub-study 2 and sub-study 3, addressing objectives 1, 2 and 3 respectively. All three sub-studies were designed to address the same overall aim (Section 1.3). Each sub-study analysed different data and has its own structure, including introduction, method, results and discussion sections. Sub-study 1 and sub-study 2 are population-based studies and analysed data supplied by the MoH. Sub-study 3, which analysed data from referrals of patients to the Waikato Hospital in Hamilton, is a part of the Health Research Council (HRC) study and the data for sub-study 3 were supplied to me by an analyst from the HRC project. The HRC study addresses the whole spectrum of the delay in CRC diagnosis; however, to the best of my knowledge, the objective for sub-study 3 was not addressed by the HRC study.

The study populations were different for each of the three sub-studies: for sub-study 1, the study population was the NZ population between 1994 and 2018, 30-<90 years old; for sub-study 2, the study population was the NZ population between 2014 and 2018, 30-<90 years old; and for sub-study 3, the study population was patients at risk of CRC who were referred to the Waikato Hospital in Hamilton between 1 January 2015 and 31 December 2017, to confirm or exclude CRC.

Patients with histologically verified incident CRC, ICD-10 codes C18-C20, were identified from the NZ Cancer Registry (NZCR), and the registration in NZCR was used as a diagnostic marker. Because the observational units are different in each sub-study, they will be explicitly stated in the method sections of the relevant chapters (Chapter 3 - for sub-study 1, Chapter 4 - for sub-study 2 and Chapter 5 - for sub-study 3).

1.4.2 Data sets

In this study, I analysed a number of comprehensive administrative data sets from the NZ health care sector available via the NZ Ministry of Health or provided by HRC. The corresponding population counts, stratified by gender and ethnicity, were sourced from Statistics NZ. Below I provide a brief description of the data sets analysed in this study.

1. Extracts from the New Zealand Cancer Registry (NZCR), including all registrations of CRC with ICD-10 codes C18–C20 diagnosed in New Zealand from 1 January 1994 to 31 December 2018 (for sub-studies 1 and 2). ICD-10 is an International Classification of Disease Coding which is used for the classification of diseases, signs, symptoms, abnormal findings, social circumstances and external causes of injury or disease. This data set includes the following information about each patient: demographics; dates of diagnosis and death; site of cancer; stage at diagnosis; and other tumour characteristics. CRC registrations made before 1 January 1994 were not included in the analysis because, prior to 1994, registrations were non-mandatory. From 1 January 1994 cancer registrations became mandatory in accordance with the Cancer Registry Act 1993 (Ministry of Health NZ, 2020b).

2. The Virtual Diabetes Register (VDR) for the years 2014–2018, which contains data about people suspected of having diabetes, identified through their use of diabetes health services (for sub-study 2). The VDR uses an algorithm to identify these people in data extracted from inpatient, outpatient, laboratory test and pharmaceutical dispensing data collections (Jo et al., 2010). The VDR does not specify the type of diabetes.

3. Hospital discharge records from the National Minimum Dataset (NMDS), for patients included in VDR years 2014–2018 (for sub-study 2). Only records which indicated *diabetes* were included. The NMDS is a national collection of public and

private hospital discharge information, including clinical information for inpatients and day patients. Private hospitals without public funding are not obliged to submit data to the NMDS.

4. The Pharmaceutical Collection (PC) data set, including records of the dispension of the medications metformin, sulfonylurea and insulin to control diabetes for the years 2013–2016, for patients included in the VDR in the years 2014–2018 (for sub-study 2). These data were available through the MoH.

5. Mortality data for the year 2018 were used to obtain information about mortality in those diagnosed with CRC in 2018; this information was not specified in the NZCR data supplied by the MoH (for sub-study 2).

6. Electronic referral (e-referral) data, consisting of all referrals made by GPs for patients with suspected CRC to the Waikato Hospital in Hamilton from 1 January 2015 to 31 December 2017 (for sub-study 3). The e-referral data set contains descriptions of symptoms which GPs considered relevant to the referral of a patient for further investigation. Information about the gastroenterologist's decision was also added to each patient's history; the data, therefore, specify whether the patient was seen or the referral was rejected, and whether or not a colonoscopy was performed. In addition, the demographics for each patient, and the information from the NZCR for each patient diagnosed with CRC during the study period, were added to the e-referral data by an HRC analyst.

7. Population count projections from Statistics NZ containing estimated population counts at 1st January and 30th June, in 1-year age groups for the entire population, for males and females, and for Māori. The tables with population count data are based on estimates for the Census years and yearly interpolation between Census years (Statistics NZ, 2020a) (for sub-study 1 and 2).

Table 1.1 specifies which data sets were used in this study to obtain the required information about patients. Data sets for specific statistical analyses were linked using encrypted patient National Health Index (NHI) numbers. The NHI system is a national index of patients that has been used in NZ since 1977 (coverage is estimated to be 98 percent of the population) (Ministry of Health, 2009). An NHI number is a unique number assigned to each patient at their first use of health services. Information about the patient is collected and maintained by GPs (through primary health care practice management systems) and through public and private hospital patient administration systems. To ensure that NHI information can be accessed only by authorised health-care professionals and pre-approved agencies as set out in Schedule 2 to the Health Information Privacy Code 1994, appropriate control and access mechanisms are applied. These include audit, security and protective mechanisms. For the protection of patients' privacy, the NHI number had been encrypted by the MoH and by HRC.

1.4.3 General approach to statistical analysis

Statistical inference in this study aims to support both doctors' decisions regarding investigation and referral, and policy-making by providing statistical models for CRC incidence rates or CRC risk in individual patients. Data will be summarised using mean and standard deviation for normally distributed variables, median and interquartile range for continuous but not normally distributed variables, and frequencies for categorical variables. 95% confidence intervals (CIs) will be provided as measures of the accuracy of the parameters of interest (Cameron et al., 2020). P-values will be provided where hypothesis testing is relevant and no meaningful effect size can be provided (e.g. the Hoshmer-Lemeshow test). P-values based on Z-test using pooled variances will be provided for the comparison of the effect sizes between different subsets or different outcomes. Interaction terms are denoted using the ":" symbol.

Information required	Data set	Source of data
Demographics*	VDR	MoH
	NMDS	MoH
	NZ Cancer Registry	MoH
CRC status	NZ Cancer Registry	MoH
Date of death	NZ Cancer Registry	MoH
	Mortality data	MoH
Diabetes status	VDR	MoH
Medication for diabetes	Pharmaceutical Collection	MoH
Symptoms and signs of CRC	E-referral	District Health Board
Family history of CRC	E-referral	District Health Board
Test results	E-referral	District Health Board
Demographics* (sub-study 3)	E-referral	District Health Board

*Demographics include: age, gender, ethnicity, deprivation

Table 1.1: *Sources of information analysed in this study.*

Standard methods commonly used in epidemiological literature, such as regression models and indicators of the diagnostic performance of symptoms and demographics, will be used where possible, to allow comparison of this study's results to results obtained from earlier studies. The methodology specific to each sub-study is explained in detail in the method section for each sub-study.

Where a large number of available covariates creates a risk of overfitting, appropriate variable selection methods will be used based on Akaike's Information Criterion (AIC) or Bayesian Information Criterion (BIC) (Dobson and Barnett, 2008) with details given in each sub-study. The following demographics were used in the statistical analyses: age, gender and ethnicity. Deprivation was used only for descriptive statistics, as Statistics NZ does not publish population tables for deprivation by age groups. Because age is the most

influential predictor of CRC, special care was taken to model age in a way that provides a good fit to the data. Age was treated as a continuous variable in all analyses.

In this thesis, the term "gender" was used throughout as opposed to "sex", as used by the NZ Census (Statistics NZ, 2018) and recommended by Statistics NZ (2020b). While, as defined by World Health Organization (WHO), "sex" refers to the classification of living things on the basis of their reproductive organs, "gender" refers to a person's self-representation as male or female or the person's role in society. Gender, although related to biology, is also influenced by a person's social environment (World Health Organization, 2021). The way how an individual perceives themselves (male or female) can relate to their health behaviour and their approach to healthcare which can be important for health research.

Stratification by ethnicity was used in some analyses of incidence data in this study. Measuring the cancer burden of Indigenous peoples is problematic, and requires careful assessment of biases inherent in cancer surveillance methods for Indigenous peoples (for a recent overview see Sarfati et al. (2018)). Appropriate care was therefore taken to understand the sources of possible bias. In all three sub-studies, ethnicity was dichotomised as Māori and non-Māori. The principle of prioritised self-identified ethnicity was used throughout, i.e. individuals who indicated multiple ethnicities including Māori were treated as Māori when assigning ethnicity to individuals based on NZCR, VDR and e-referrals.

For tabulations from Statistics NZ, this dichotomisation was achieved by using the Māori tabulations as provided, while calculating the non-Māori population counts by subtracting the number of Māori from the total population count (for each year–age combination). Statistics NZ has been using self-identified ethnicity since 1986, i.e., from before the study period.

For NZCR data, the quality of the ethnicity information is reviewed in detail in Shaw

et al. (2009). Ethnicity is in principle recorded in a way similar to the Census, but since ethnicity is sometimes missing in the recording forms, NZCR uses other sources (NMDS, NHI and the mortality register) to populate the ethnicity field. According to the review, the number of Māori in NZCR was undercounted relative to the Census by approximately 20% in 1994, improving to 15% in 2004. According to Boyd et al. (2016), there has not been systematic under-counting of Māori in NZCR since 2006.

For VDR, prioritised ethnicity is obtained from the NHI system by the MoH. During literature searches I have not found any review of the quality of the ethnicity data in NHI. However, Boyd et al. (2016) reported good quality ethnicity records in NZCR and the mortality register; the VDR relies partially on both, and I therefore assumed that the quality of the ethnicity information in the VDR 2014–2018 is reasonable.

1.4.4 Research strategy

The strategy of this study fits into the Model of Pathways to Treatment proposed by Walter et al. (2012), based on the seminal work by Andersen et al. (1995). The model provides a framework for reducing the delay in cancer diagnosis and time to treatment including the following intervals within which delay in cancer diagnosis can occur: symptom appraisal by the patient; help seeking; diagnostic interval and pre-treatment. This study addresses only the diagnostic interval. The diagnostic interval includes several steps from the first consultation with a health provider to the diagnosis, explained in a later section (Figure 2.2 in Section 2.4.2).

Figure 1.2 provides a summary of the research strategy which links the study objectives to the overall aim i.e., addressing the CRC burden (green boxes). The figure also shows other possible research paths to improving CRC outcomes that are not addressed in this study (grey boxes).

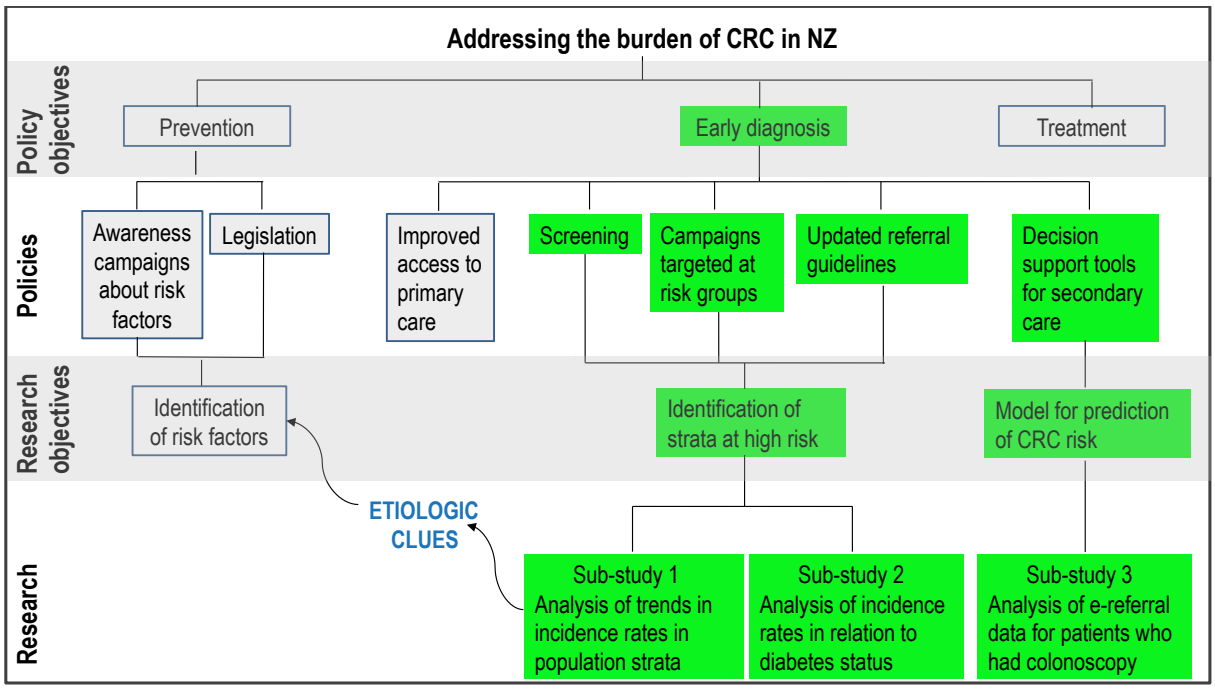


Figure 1.2: *The strategy for the study. The objectives addressed and tasks undertaken in this study are indicated in green rectangles.*

1.5 Ethical approval and data management

This study was first approved by the Central Health and Disability Ethics Committee (HDEC) on 16 July 2018 (approval number: 18/CEN/118). Additional approval was subsequently obtained on 13 September 2018 (approval number: 18/CEN/118/AM01). Letters of approval can be found in Appendix A, B, C and D.

The Human Research Ethics Committee was informed about the HDEC approval. The study was endorsed by Te Puna Oranga Māori Consultation Research Review Committee of the District Health Board (DHB) on 2nd July 2018. The Te Puna Oranga letter is included in Appendix E.

Sub-study 3 was carried out under the ethical approval of the HRC study (approval number: 17/417, "Reducing delay and increasing access to early diagnosis for colorectal cancer").

Although the data were supplied using the encrypted NHI, the data include sensitive and personal information about patients. The data were therefore managed in accordance with the Privacy Act 1993, and the specific Codes of Practice adopted by the Privacy Commissioner. The data were supplied by the MoH on an encrypted memory stick, with the password sent in a separate email and known only to Malgorzata Hirsz. The original data from the MoH were stored on University of Waikato servers and only the following authorised people had access to the data: Malgorzata Hirsz, Lyn Hunt, Lynne Chepulis, and Michael Mayo. The principal researcher's version of the data, including processed anonymised data used in analyses, was stored in an encrypted personal file with a secure password. The data will be removed from the server and from the researcher's personal computer no later than the completion of the study.

1.6 Thesis structure

This section presents the structure of the thesis. The thesis is divided into seven chapters. Chapter 1 provides the rationale for the research, discusses research objectives, outlines the study design, and presents the thesis structure. Chapter 2 summarises the background knowledge that supports this study.

The subsequent three chapters present the empirical components of the study, addressing the three objectives stated in Section 1.3. Chapter 3 presents sub-study 1, which analysed long-term trends in CRC incidence. Chapter 4 presents sub-study 2, which estimated the association between diabetes and CRC. Chapter 5 presents sub-study 3, which developed a predictive model for CRC risk in the population referred by GPs to the secondary care enabling discrimination of patients with low CRC risk during the triage process.

Finally, Chapter (6) provides a synthesis and overall discussion of the findings from all three sub-studies, with an emphasis on the novel results the study adds to existing knowl-

edge. The chapter discusses examples of the possible use of the study's results in clinical practice, discusses the limitations and strengths of the study, gives the direction of the future research that emerged from this study, and finishes with short closing remarks.

1.7 Summary

Chapter 1 outlined the whole project and provided an explanation of the problem to address. In this chapter, I also gave the rationale for the research; I explained the methodology for the entire study; and outlined the structure of the thesis to give the roadmap for the content of the following chapters. The background knowledge and literature review which led to the identification of the gaps which the study aimed to address are presented in the next chapter.

Chapter 2

Background

2.1 Introduction

This chapter provides a brief introduction to CRC and presents the research that provides the evidence-base for the rationale of the study. The following sections discuss: the clinical description of CRC; risk factors in CRC; the signs and symptoms present in patients with CRC; how CRC is diagnosed; a summary of the epidemiology of CRC, including incidence, survival, non-modifiable risk factors in CRC, and examples of initiatives undertaken to decrease the burden of CRC.

The analysis of studies conducted elsewhere sought to identify research which can help to tackle the burden of CRC using methods that could be carried out during the three-year study and could give sufficiently accurate estimates (with narrow 95% CIs) for results to be useful in clinical practice. The review includes research on risk factors associated with CRC that could be used as additional predictors in clinical practice and are available to clinicians without performing expensive or invasive tests. It also examines CRC surveillance research which uses population-based incidence data and can provide estimates applicable to the whole population. Such estimates can, in turn, be used in primary care, in revision and updating the policy for population-based screening, and also for generating hypotheses about etiological factors. Finally, the review presents predictive models designed for use as supportive tools for secondary care specialists.

This review does not include studies investigating genetic- and bio-markers for earlier detection of patients with CRC, and studies on post-diagnostic care.

The searches were limited to Google Scholar to identify articles on colorectal cancer. This choice was made because Google Scholar indexes, not only academic papers, but also NZ Government reports and non-peer-reviewed articles. Publications were limited to human studies published up to December 2020 in English and Polish languages. The search strategy included combinations of the terms: colorectal cancer, cancer, colon, rectum,

proximal, distal, risk factor, symptoms, diagnosis, incidence, survival, cancer surveillance, screening, age-period-cohort, predictive model, diabetes, New Zealand. Additional literature was also identified from references included in selected papers.

2.2 Basic physiology of the large intestine and colorectal cancer

Colorectal cancer includes cancerous growths in the colon, rectum and appendix, collectively called bowel, colorectum, or large intestine (Figure 2.1).

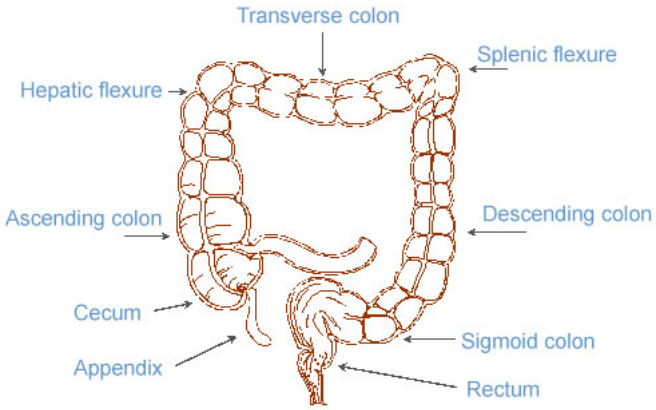


Figure 2.1: *Anatomy of the large intestine. Source: National Cancer Institute, https://training.seer.cancer.gov/colorectal/anatomy/figure/figure2.html*

The colon is a muscular tube about 1.5 meters long and 5 centimetres in diameter and is divided into four sections: the ascending colon that begins with the cecum and extends upward on the right side of the abdomen; the transverse colon which crosses the body from the right to the left side; the descending colon which descends on the left side; and the sigmoid colon, which is the final portion of the colon and joins the colon with the rectum. The ascending and transverse colon are collectively referred to as the proximal colon, while the descending and sigmoid colon are collectively referred to as the distal

colon. The rectum is the last 15 cm of the large intestine.

Proximal and distal colon differ with respect to the embryological origin. Relative to the splenic flexure, the proximal part of the colon originates from the midgut, while the distal colon and the rectum originate from the hindgut (Bufill, 1990). The distinction according to the embryological origin is important due to the differences in colorectal tumours, with their clinicopathologic characteristics depending on the distal or proximal location (Bufill, 1990). This difference led Bufill to hypothesise that the proximal and distal parts of the colorectum might also differ with respect to the susceptibility to several risk factors, e.g. to increased level of indigenous insulin (hyperinsulinemia) or to increased level of blood sugar (hyperglycemia). The distinction is also relevant to procedures used to diagnose CRC: colonoscopy can visualise the whole colorectum, while flexible sigmoidoscopy can visualise only the proximal part of the colorectum. The diagnostic procedures are explained in more detail in Section 2.4.

Most colorectal cancers develop from adenomatous polyps which represent around half to two thirds of all colorectal polyps (Levin et al., 2008). In most patients this is believed to take 10–20 years which gives an opportunity for effective intervention i.e. removal of polyps before progression to cancer (Half and Arber, 2009).

2.3 Symptoms of colorectal cancer

CRC may manifest itself with several symptoms including: rectal bleeding, abdominal pain, abdominal mass, constipation, unexplained weight loss, change in bowel habit, and anaemia, with some features present more often than others and possessing different diagnostic values (Jellema et al., 2010). A UK study by Hamilton et al. (2005) reported rectal bleeding, weight loss, prolonged duration of abdominal pain, constipation and diarrhoea as independent predictors associated with diagnosis of colorectal cancer. A

systematic review of 23 studies conducted by Astin et al. (2011) investigated symptoms of colorectal cancer in the primary care population. It reported rectal bleeding as the most predictive symptom for CRC among investigated symptoms.

Anatomical location of tumours has been reported to influence the clinical manifestation of CRC. For example, the Polish study by Banaszkiewicz et al. (2009) reported abdominal pain, anaemia and tiredness, but not constipation, to be commonly present in patients with proximal tumours, and rectal bleeding and constipation often presented in distal tumours. In a US study conducted by Majumdar et al. (1999), patients with distal tumours similarly presented with rectal bleeding and constipation and patients with proximal tumours often experienced symptoms such as loss of appetite, nausea, vomiting, abdominal pain and tiredness, but in addition Majumdar also reported change in bowel habits and anaemia as common symptoms in patients with tumours located in both sides of the colorectum. However, iron deficiency anaemia was found to be much more common in patients with proximal tumours and change in bowel habits more common in patients with cancers located in distal colon in another study conducted in Iceland (Alexiusdottir et al., 2012).

The differences reported could be due to different study designs rather than due to the real differences in CRC presentation in different populations. Although Majumdar et al. (1999), Banaszkiewicz et al. (2009), and Alexiusdottir et al. (2012) all used a retrospective design, the studies differ with respect to inclusion criteria and where the symptoms were elucidated from; Majumdar et al. (1999) and Banaszkiewicz et al. (2009) analysed data only from symptomatic patients, while Alexiusdottir et al. (2012) included symptomatic and asymptomatic patients. Further, Majumdar et al. (1999) used primary care and hospital notes to extract symptoms, while Alexiusdottir et al. (2012) and Banaszkiewicz et al. (2009) used only hospital notes.

With respect to symptoms associated with rectal cancer, rectal bleeding and diarrhoea

have been reported as often occurring characteristics (Banaszkiewicz et al., 2009), however, in general symptoms of rectal cancer are non-specific and are similar to symptoms presented in cancers located in distal colon (Gaertner et al., 2015).

2.4 Diagnosis of colorectal cancer

CRC can be diagnosed in one of three ways: during a screening program as an asymptomatic patient; during investigation for possible CRC as a symptomatic patient; or during investigation or treatment for other disease, e.g. appendicitis (Ewing, 2018). The latter possibility is not addressed in this thesis. Timely diagnosis is always the goal, regardless of the mode of investigation. If our fictional patient Mary from Figure 1.1 has already existing and diagnosable CRC, the goal of the medical care is to diagnose her as soon as possible to avoid unnecessary delay. The importance of avoiding delay in cancer diagnosis is described in *The New Zealand Cancer Control Strategy* document (Ministry of Health and the NZ Cancer Control Trust, 2003):

> The diagnosis of cancer covers a breadth of activity, from presentation or identification of signs and symptoms, to confirmation (or elimination) of a cancer diagnosis. For those with cancer, their family and whānau, the definitive diagnosis of cancer is the beginning of a journey, the duration of which can extend from months, to years, to a lifetime. Of prime importance is the timeliness of diagnosis. An excessive delay between the presentation or identification of initial symptoms and the definitive diagnosis can have a significant psychological effect on those with cancer, their family and whānau. This, along with a further delay to definitive treatment, can have an impact on the likely effectiveness of treatment.

The remainder of this section gives a brief explanation of methods used to diagnose CRC

and an overview of the diagnostic pathways for asymptomatic and symptomatic patients in NZ.

To date, colonoscopy is the gold standard in CRC diagnosis. Although colonoscopy has the ability to identify patients in very early stages, it is an invasive test with a non-trivial risk of serious health complications (Ransohoff, 2005). Moreover, in NZ, similarly like in many other countries, the capacity for performing colonoscopies is limited and, therefore, too many referrals of patients with low CRC risk can create a long waiting list and result in delayed diagnosis of many patients who actually have underlying cancer (Yeoman and Parry, 2010; Stamm et al., 2020). It is therefore important to select for colonoscopy those with sufficiently high risk of CRC.

There are other methods that can be used to diagnose CRC such as flexible sigmoidoscopy (FS), computer tomography and barium enema. For an overview of available methods, see Kolligs (2016). FS, which visualises only the distal part of the colorectum, is of particular interest of this study as patients with high risk of distal tumours and low risk of proximal tumours can be initially investigated using FS which is a less invasive investigation then a colonoscopy and therefore FS could be the preferred investigation for some patients. Also, replacing the initial investigation with colonoscopy in some patients by FS could help in better management of the existing colonoscopy resources.

2.4.1 Investigation of asymptomatic patients

National bowel screening programs are strategies designed for early detection of CRC in asymptomatic patients. Screening programs have already been implemented and evaluated in developed countries such as Australia, Canada, Norway, US and UK (Zauber, 2015). Screening for CRC has been shown to decrease rates of diagnosis in late stages by finding early stage cancers and cancer precursors in asymptomatic populations, which improves the chances for complete resection of the lesions (Nielsen et al., 2011; Zauber,

2015).

In NZ, the National Bowel Screening Programme (NBSP) has been initiated gradually from 2017, and is expected to be implemented in the whole country in 2021 (Ministry of Health NZ, 2017). The screening in NZ, based on biannual immuno-histochemical faecal occult blood test (FOBT), has been offered to asymptomatic individuals at age 60–74, with the following investigation (mostly with colonoscopy) for those with a positive FOBT result. Once the screening program has been fully introduced, there is a concern that NZ might not be able to deal with the increased number of colonoscopies needed. As a result, the diagnostic pathway for symptomatic patients might be impaired, leading to further delay in the diagnosis of symptomatic patients (Yeoman and Parry, 2010; Stamm et al., 2020).

As an alternative solution, population-based screening in NZ could be based on FS, as Cox (2016) has suggested. Cox (2016) argues that FS has been shown to have higher effectiveness in preventing future occurrence of CRC, as well as deaths from CRC, compared to the screening modality based on FOBT. Although Cox's (2016) argument is convincing, the screening program which was introduced in NZ is based on FOBT, with patients with a positive FOBT test having a colonoscopy.

A key focus of the NBSP is equity with respect to health gains for Māori and non-Māori (Ministry of Health NZ, 2018b). However, according to calculations by the MoH, the NBSP might increase inequities. The MoH considered lowering the screening age for Māori to 50 years but, in line with recommendations from the Bowel Screening Advisory Group (BSAG), the MoH decided to monitor the situation for the next three to four years before reconsidering the proposal. It is necessary to keep in mind that, as Richardson and Potter (2014) explain, no screening test is perfect, and screening can do harm, even assuming that a screening programme is beneficial. It is therefore not obvious that it will be in the best interests of Māori to initiate screening at age 50. One of the reasons

given by the MoH for lower benefits from screening in Māori, was the comparatively lower risk of CRC in Māori (Ministry of Health NZ, 2018b). However, as previously discussed, CRC incidence rates are converging in Māori and non-Māori. At the same time, published research on age-specific CRC rates by ethnicity in NZ is scarce. It is therefore not known whether Māori actually are currently at lower risk of CRC at screening age. Knowledge about age-specific incidence rates by ethnicity would thus be very useful for policy-makers when considering future changes to the screening policy.

Data from other countries that have already implemented screening report variable rates of participation in screening programs, with only around 23% of the eligible population participating in Australia, compared to 56% in the UK and 81% in Japan (Khalid-de Bakker et al., 2011). In NZ, a pilot study of attendance at two rounds of screening in Waitematā estimated that 56.9% of eligible individuals took part in the first round of the screening program (Ministry of Health NZ, 2016). Māori (with 46% participation) were less likely to participate in the screening than non-Māori. Additionally, over 20% of CRC patients in NZ are diagnosed before the age of 60 (23.5% of patients in 2017 (Ministry of Health NZ, 2019d)). In this situation, for many NZ patients, early diagnosis depends on the diagnostic route through primary care, which is explained in the next section.

2.4.2 Investigation of symptomatic patients

Investigation of symptomatic patients in NZ mostly follows a diagnostic pathway via primary care shown in Figure 2.2. The diagnostic pathway starts from the time when a patient detects the first symptoms and decides to make an appointment with the GP. If the GP suspects the patient to be at high risk of CRC, the patient can be referred directly for a colonoscopy or to a specialist for the first specialist assessment (FSA). The referral for FSA can be accepted or rejected by the specialists. Most patients with accepted referrals will have a colonoscopy performed. Patients with rejected referrals will return

to GPs for further surveillance.

This mode of diagnosis does not apply to individuals diagnosed after presentation in ED. Diagnosis of CRC after initial presentation of a patient in ED falls outside of the scope of this thesis. The diagnostic pathway as illustrated in Figure 2.2 is simplified; in reality patients often visit GPs multiple times and might see other health care providers before visiting a GP (Windner et al., 2018). Windner et al. (2018) constructed a more complex pathway, however based on a small sample of 98 patients, and the authors acknowledged, that the sample was not representative of the New Zealand population diagnosed with CRC and might not fully reflect the real pathway.

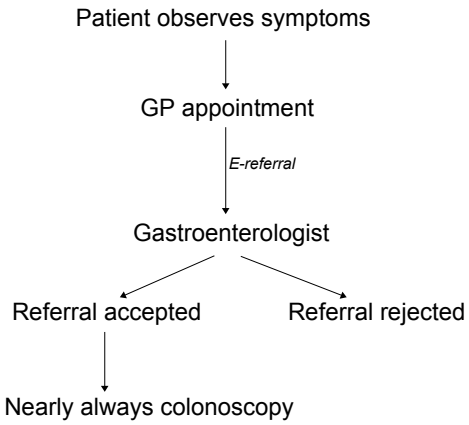


Figure 2.2: *Diagnostic pathway for CRC diagnosis in NZ.*

2.4.2.1 Investigation in primary care

In NZ, as in other countries with the universal first contact via general practice, GPs play a crucial role in the early diagnosis of CRC (Fletcher, 2009). GPs can facilitate early diagnosis through a prompt referral of patients with suspected cancer for further investigation, either by a direct referral for a colonoscopy or through a referral for the first specialist assessment (FSA) (Ministry of Health NZ, 2020c). It is desirable that GPs

choose appropriate patients, i.e. those who are more likely to have underlying, but not yet diagnosed, CRC. The choice is problematic, and it is seen as a multifactorial problem (Langenbach et al., 2003; Fletcher, 2009; Ewing, 2018). The process of CRC diagnosis is very difficult compared to diagnosis of e.g. skin cancer or breast cancer, because of the location of the tumours which requires invasive investigation using colonoscopy to rule in/out cancer (Fletcher, 2009; Hamilton et al., 2009).

The majority of symptoms which CRC patients present to a GP are not specific for CRC and are often also associated with a non-malignant disease such as irritable bowel syndrome (IBS) or inflammatory bowel disease (IBD) (Jellema et al., 2010; Lyratzopoulos et al., 2014). Those symptoms include weight loss, changes in bowel habit, constipation, diarrhoea, tiredness, bleeding or abdominal pain. In primary care, such symptoms account for up to 10% of all consultations (Jones, 2008). This further complicates the identification of patients who are at sufficiently high risk of CRC to justify referral to the specialists or a direct colonoscopy. As well, the fact that, on average, a GP sees only one new patients with CRC per year who presents with one or more of those unspecific symptoms (Hamilton and Sharp, 2004) makes it difficult for GPs to build up experience in recognising which patients are more likely to have CRC.

During assessment, GPs in NZ use referral guidelines for suspected CRC developed by the NZ Guidelines Group (New Zealand Guidelines Group, 2009). The NZ guidelines are based on the National Institute for Clinical Excellence (NICE) guidelines (National Institute for Health and Clinical Excellence, 2005), updated in 2015 and 2017 (National Collaborating Centre for Cancer, 2015). The NICE guidelines include many, but not all, of the symptoms which might be present in CRC patients (Jellema et al., 2010). The current NICE guidelines classify well those patients who are at high risk of CRC and present with already advanced cancer (Hippisley-Cox and Coupland, 2012). The guidelines fail, however, to identify a significant number of patients who might have cancer in earlier, more curable stage (Jones et al., 2001). Patients with so-called "red flag" symptoms such

as rectal bleeding, unexplained weight loss or abdominal mass have priority for urgent further investigation. However, according to Jones et al. (2007), the approach based on "red flag" symptoms could identify only between 20 and 40% of patients who actually had CRC and had any symptoms within three years prior to diagnosis. This implies that most patients who in fact had CRC and presented to GPs had not so severe symptoms. In addition, the majority of patients with "red flag" symptoms will not have cancer, as the positive predictive values (PPV) of such symptoms are <10% in males (Jones et al., 2007), around twice as low for females and even lower for young adults (<1%) (Dommett et al., 2013).

The NZ Guidelines Group included many of the NICE guidelines recommendations with only minor changes related to investigation of patients with family history of CRC and management of patients with inflammatory bowel disease (New Zealand Guidelines Group, 2009). Despite the referral guidelines, it is difficult for GPs to select the right patients for the appropriate diagnostic pathway for all those who present with symptoms. To address the difficulties with identification of patients at risk in primary care, and to assist GPs during assessment of a patient, Marcela Ewing, a Swedish medical doctor with experience as a GP, wrote her PhD thesis addressing the "Identification and early detection of cancer patients in primary care" (Ewing, 2018). Ewing aimed to determine whether it is possible to detect cancer earlier, at a less advanced stage. For the selection of patients with non-metastatic CRC in primary care, Ewing and colleagues (Ewing et al., 2016) proposed a tool based on demographics and a combination of symptoms. Similar algorithms have been proposed based on data from different countries, e.g the UK (Hippisley-Cox and Coupland, 2013) and Israel (Kinar et al., 2016).

Multivariable prediction models that use a combination of risk factors for the calculation of CRC risk, but without concentrating on diagnoses of early stages (Hamilton, 2009b; Marshall et al., 2011; Hippisley-Cox and Coupland, 2012), have also been shown to improve performance when compared to single symptoms. Such models could help GPs to

select appropriate patients for further investigation. Such research, however, was possible only due to the high quality of electronic patient records from the primary care settings available in e.g. Scandinavia, Israel and the UK. Because NZ is lacking such a database, models of this kind cannot be derived from data obtained from New Zealanders. Calibration of one of the already-existing models would be potentially possible, but it would require strong assumptions that might not hold true in real life: for example, the prevalence of symptoms may not be the same in the NZ population as in the population from which the model was derived.

In NZ, GPs have an option to refer patients with strong suspicion of CRC directly for urgent colonoscopy, bypassing the FSA, using a different set of guidelines which are summarised in Ministry of Health NZ (2019c). However, symptoms and patient characteristics (age eligibility for referral) included in these guidelines are limited and patients presenting with symptoms not specified in those guidelines would not qualify for urgent referral. As a result, if GPs always adhere strictly to guidelines, many young patients and older patients with atypical presentations would have delayed investigation and diagnosis.

Importantly, as stated in National Collaborating Centre for Cancer (2015), guidelines and recommendations are not requirements, and the important role of doctors' judgement is highlighted by the document. "Recommendations do not override clinical judgement. It is well recognised that primary care clinicians have expertise in recognising patients who are "ill" and in knowing that "something is wrong". NICE guidelines have supported the idea that clinical intuition has diagnostic value: "..... clinicians should trust their clinical experience where there are particular reasons that this guidance does not pertain to the specific presentation of the patient" (National Collaborating Centre for Cancer, 2015).

To summarise, the use of the guidelines alone is not always sufficient to decide whether a patient has a sufficiently high CRC risk to justify a referral for FSA or direct referral

for colonoscopy. Other methods for decision-making are used, including so called pattern recognition which is doctors' intuition based on previous experience with similar patient (Sackett et al., 1985; Elstein and Schwarz, 2002). There are also statistical models and algorithms that doctors can use for making decisions (Williams et al., 2016). Algorithms, like guidelines, are not meant to replace doctors' judgement, but rather provide an aid in the assessment of patients which may be especially useful for the assessment of borderline cases.

2.4.2.2 Investigation in secondary care

In NZ, for patients suspected of CRC in primary care, the referral can be made to the gastroenterology and general surgery departments in the hospital for FSA. The referral can be accepted or rejected by a specialist (Leaman, Aaron , 2017). The decision is based on the information specified by GPs in the referrals, sometimes very limited, as well as on the patient's demographics. Additionally, there might be other factors which physicians take under consideration from e.g., conversation via email with the GP. Due to the limited resources for colonoscopies and gastroenterology services in NZ (Stamm et al., 2020), some referrals are rejected (Leaman, Aaron , 2017).

As a colonoscopy is a limited resource also in other countries, Adelstein et al. (2010) presented a model based on Australian data for use in secondary care settings to select those with high enough risk of CRC to justify a colonoscopy. The proposed model could successfully select 95% of CRC patients with only 60% of the colonoscopies actually performed. This indicates that a large proportion of patients with low CRC risk are undergoing a colonoscopy, an invasive procedure which could be avoided in some patients with low CRC risk. The idea explored by Adelstein et al. (2010) seems to be an attractive option to help with management of the limited gastroenterology resources in NZ. Such a model designed using NZ data, to the best of my knowledge, has not been published

yet.

2.5 Epidemiology of colorectal cancer

The purpose of this section is to describe the burden of colorectal cancer in NZ, and also to show some existing gaps in research related to the epidemiology of CRC in NZ.

2.5.1 Incidence of colorectal cancer

Colorectal cancer (CRC) is a big burden in the developed world (Favoriti et al., 2016; Keum and Giovannucci, 2019), with incidence and mortality rates varying considerably between countries (Torre et al., 2016; Gandomani et al., 2017; International Agency for Research on Cancer, 2020; Keum and Giovannucci, 2019). The incidence rates of CRC in NZ are among the highest in the world (Figure 2.3) (International Agency for Research on Cancer, 2020). In NZ, the incidence rates for CRC in 2018, age-standardised to the structure of the WHO Standard Population (ASR), were 35.3 per 100,000 for all, 40.3 per 100,000 for males and 30.8 per 100,000 for females (International Agency for Research on Cancer, 2020).

In New Zealand, CRC is the second most common cancer in males and females after lung and breast cancer respectively, and the second-leading cause of cancer-related death (Firth et al., 2016; Ministry of Health NZ, 2019b). About 3000 new patients are diagnosed with CRC each year, and about 1200 patients per year die from CRC, which, in 2017, accounted for around 12% of all cancer related deaths (Ministry of Health NZ, 2019a). Because of the strong positive association of CRC with age, it is expected that the incidence rates will increase worldwide, including in NZ, as the population is ageing (Douaiher et al., 2017). Prevention is very important, but it is unlikely that even very

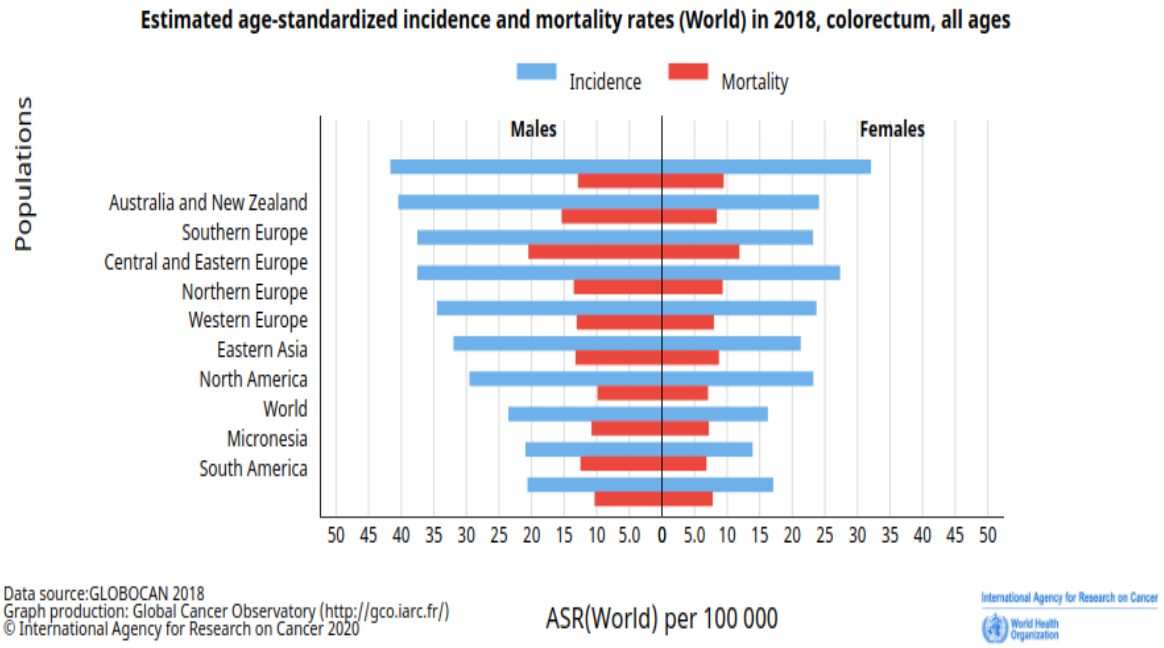


Figure 2.3: *Age-standardised CRC incidence and mortality rates in different parts of the world (International Agency for Research on Cancer, 2020).*

well implemented health promotions will reverse the growing incidence rates due to the ageing of the population. Similarly, advances in medical treatments will not contribute greatly to the improvement in survival rates in patients diagnosed with late stage of CRC (Thorne et al., 2006). Therefore, improvement of early diagnosis is a universally agreed strategy for improving long-term survival in CRC (Hamilton, 2009a).

2.5.2 Trends in incidence in NZ

Although this study investigates trends in CRC incidence between 1994 and 2018 only, i.e., since cancer registrations in NZCR became compulsory, I briefly summarise the historical incidence trends from year 1948 (when NZCR was established) (Ministry of Health NZ, 2020a). The ASR increased steadily up to year 1994 and since then has been decreasing, similarly in males and females (Figure 2.4). The increase in ASRs may be real but most likely also reflects the increasing coverage of CRC diagnoses in the

NZCR. After 1994, the decrease would be expected to reflect a real decrease in ASRs as the coverage of CRC diagnoses in the NZCR is reported to be close to complete (Cunningham et al., 2008).

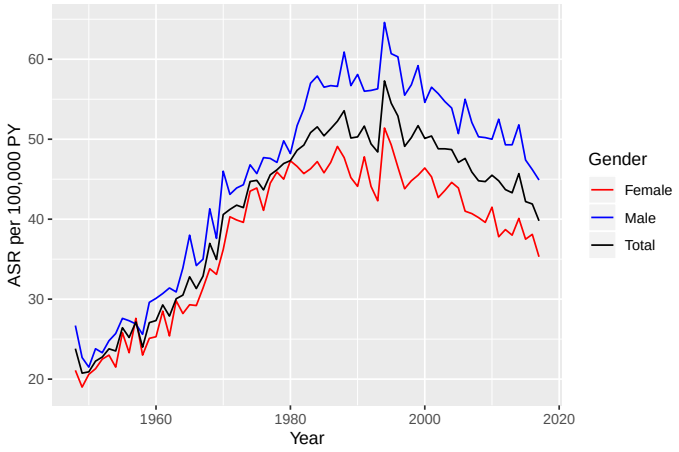


Figure 2.4: *CRC incidence rates in New Zealand, 1948–2017, age-standardised to the WHO 2000 population (Ministry of Health NZ, 2020a).*

The slowly and steadily declining trend in ASRs in NZ from around year 1995 have also been reported in literature which compared data from multiple countries (Arnold et al., 2017; Siegel et al., 2019; Araghi et al., 2019). Arnold et al., who investigated associations between CRC incidence trends and the Human Development Index (HDI), a score that reflects the economic development of a country (Stanton, 2007), reported that NZ followed a general pattern in CRC incidence trends seen in countries with very high HDI such as Japan, USA, Iceland and France. In those countries, ASRs have been declining over recent decades, presumably in response to improved prevention and uptake of CRC screening (Siegel et al., 2014; Zauber, 2015; Siegel et al., 2019). However, in NZ the decrease cannot be attributed to screening. A better understanding of the decline in CRC incidence in NZ, a country which did not have population-based CRC screening at that time (the NZ incidence data analysed by Arnold et al. (2017) included years 1984–2010), would therefore be of interest, not only for the NZ health sector and researchers,

but also for the international epidemiological community.

Regardless of the overall decrease in ASRs in New Zealanders showed by Arnold et al. (2017), a later study by Siegel et al. (2019) who analysed CRC incidence data from 36 countries including data from NZ for period 2007–2016, reported that ASRs in NZ decreased only in individuals older than 50 years, while in those between 20 and 50 years old ASR increased substantially, with similar trends also observed for Australia, Canada, US and Germany. The increasing rates in young New Zealanders have been reported in both genders for colon and rectal tumours (Siegel et al., 2019; Keum and Giovannucci, 2019). Another study by Araghi et al. (2019) which analysed incidence data from seven high-income countries, also reported increasing incidence in young New Zealanders for both colon and rectal tumours. However, a NZ study by Gandhi et al. (2017) reported the increased incidence in individuals younger than 50 years, but only in rectal cancers.

Both studies, Keum and Giovannucci (2019) and Araghi et al. (2019), analysed the same NZ data; however, while Keum and Giovannucci (2019) presented only ASR, Araghi and colleagues used age-period-cohort analysis. Results of an age-period-cohort analysis are the most informative when it comes to trend analysis. Although the trend analysis performed by Araghi et al. (2019) presented a very elaborate comparison between incidence trends in the seven high-income countries, there is no evidence provided that goodness of fit analysis was performed and that the model was evaluated for the fit of the NZ data. The analysis used age categorised in 5- or 10-year age groups (depending on the sub-analysis), and only used data up to the year 2014. Due to the availability of newer data and the possibility to use a model specifically validated for fitting the NZ data, with a 1-year age resolution, it is important to provide updated estimates.

The trends in incidence rates in NZ have been shown to differ between Māori and non-Māori (Shah et al., 2012; Teng et al., 2016). The incidence rates have historically been reported to be lower in Māori than in non-Māori but between 1981 and 2004 (in Shah

et al. (2012)), or between 1981and 2011 (in Teng et al. (2016)), CRC incidence has been reported by both authors to be increasing in Māori, reducing the gap in incidence between those two ethnicities. However, both authors included in their analyses early data from NZCR from before year 2006, when many Māori CRC patients were recorded as non-Māori in NZCR (Shaw et al., 2009). This undercount of Māori ethnicity in NZCR could lead to inaccurate estimates of the trends as the degree of undercounting varies between years (Shaw et al., 2009; Boyd et al., 2016).

Shah et al. and Teng et al. both used probabilistic linkage of NZCR records to census records, and therefore the ethnicity of CRC patients in their studies was assigned based on ethnicity recorded in the census. However, as ethnicity was one of the linkage criteria, those Māori CRC patients who had non-Māori ethnicity recorded in NZCR had big probability of not being linked. Both authors subsequently weighted NZCR records based on the inverse linkage probability which in this case could lead to an inflated number of non-Māori instead of correcting for the undercount of Māori in the NZCR. This problem was explained by Shaw et al. (2009) who published correction factors for the undercount of Māori CRC patients in the NZCR. Shaw et al. did not use ethnicity as a linkage factor for the estimation of the correction factors.

With respect to the statistical methods for investigation of CRC trends used in earlier studies, most papers provide the value of the changes over a specific period as a result based on join-point analysis, a widely used technique for providing the average annual percentage change (AAPC) in incidence rates. In order to better understand long-term time trends in incidence and mortality in large populations, advanced statistical methods, known as age-period-cohort (APC) models, have been developed Holford (1983); Clayton and Schifflers (1987a) and encouraged for use as a preferable method for analysis of trends in cancer registry data (Rosenberg and Anderson, 2011; Smith et al., 2016). In NZ, APC models have been used by the MoH for forecasting purposes (Ministry of Health NZ, 2010) but I have not found any interpretation of the age, period and cohort effects.

2.5.3 Survival in colorectal cancer

CRC post-diagnosis survival rates in NZ patients are low compared to the rates in similarly developed countries (Aye et al., 2014; Elwood et al., 2016; Arnold et al., 2019). In a report of the overall 5-year survival in seven high-income countries in years 1995–2014, survival in NZ patients from colon and rectal cancers was shown to be lower than in their Australian, Canadian and Scandinavian counterparts (Arnold et al., 2019). These authors found that in NZ 5-year survival rates showed little improvement between the years 1995 and 2014. The 5-year survival in NZ was 64.5% for colon and 69.3% for rectal cancer compared to e.g. Australia with 73.3% and 75.0% (in the last analysed 5-year period 2010–2014).

The stage of the tumour at diagnosis is considered the most important predictor of survival (Hamilton, 2009a; Vega et al., 2015). The US study by O'Connell et al. (2004) reported 5-year survival in patients with stage I to be 93.2% and with stage IV only 8.1%, while a NZ study by Buchwald et al. (2018) reported 5-year survival in stage I to be 80% and in stage IV, 14% (the estimate is based on data from Christchurch Hospital years 1993–2009). A German study by Majek et al. (2012) including nearly 165 000 patients reported 5-year relative survival for localised colon cancer to be 91.3% and for metastatic stage 15.3%; and for rectal cancer 85.9% for localised and 14.2% for metastatic stage. For comparison, in NZ, the 5-year survival rates in those diagnosed in 2007 and 2008 were lower; for colon cancer in stage I only 80% and in stage IV 6%; for localised rectal cancer only 65% and 10% for metastatic (Sharples et al., 2018). As these numbers show, NZ patients have poor survival compared to patients in some other countries regardless of the stage at diagnosis. Those stage-specific discrepancies are likely to be due to differences in post-diagnostic care which is not the focus of this study.

On the other hand, the poor survival rates in NZ patients are also partially attributed to the disease being diagnosed at advanced stages more often than in other developed

countries (Samson et al., 2009). As suggested in earlier research, increased awareness of symptoms among patients and earlier diagnosis alone could improve survival, even without any new medical advances (Richards, 2009; Hippisley-Cox and Coupland, 2012). In NZ, an estimated number of 600 deaths from CRC between 2006 and 2010 could be avoided if patients with CRC were diagnosed earlier, i.e., with a stage distribution comparable to that in Australia (Sandiford et al., 2015).

One factor which contributes to the high number of diagnoses at late stages is a large proportion of New Zealanders diagnosed after presentation with symptoms in the Emergency Departments (ED), as presentation in ED is associated with more advanced disease (McArdle and Hole, 2004). Sharples et al. (2018) reported that around 31% of colon cancers in NZ were diagnosed in the ED, compared to 21% in the UK (Scott et al., 2013). Within NZ, Māori are more likely than non-Māori to present with CRC in ED (Hill et al., 2010; Sharples et al., 2018).

Survival rates in CRC patients is not a scope of this study, however, I provided this information to show that given the strong association between stage at diagnosis and survival, improvement of early diagnosis is one of the ways to achieve better outcomes in NZ patients.

2.5.4 Risk factors for colorectal cancer

An individual's risk of developing CRC depends on many factors that, for simplicity, are often divided into two categories: modifiable and non-modifiable risk factors (Haggar and Boushey, 2009; Johnson et al., 2013). **Modifiable risk factors** include factors related to lifestyle such as: obesity; sedentary life-style; lack of physical activity; nutrition; high consumption of red and/or processed meat; diet low in dietary fibre; excessive drinking of alcohol; and smoking (Burkitt, 1971; Giovannucci, 2002). A significant part of CRC risk could be attributable to such lifestyle factors (Platz et al., 2000; Chan and Giovannucci,

2010). Those factors, however, are not the concern of this study and will not be discussed further.

Non-modifiable risk factors include: increasing age; gender; ethnicity; family history of CRC; and inflammatory bowel disease (IBD). CRC has also been reported to be associated with other health conditions (Gonzalez et al., 2001; Rawla et al., 2019), among which diabetes mellitus is in interest of this study due to the sound evidence for the association reported in earlier research (Larsson et al., 2005; Tsilidis et al., 2015), due to the high prevalence of diabetes mellitus in the NZ population, and the availability of a population-based registry of patients with diabetes (VDR). Since non-modifiable risk factors and diabetes mellitus are the main focus of this study, a detailed review is provided in the remainder of this section.

2.5.4.1 Age

Age is the most important risk factor for CRC. Although the disease can develop at any age, it is diagnosed mostly in older people, and the incidence rises sharply in people older than 50 years (Haggar and Boushey, 2009), with over 70% occurring in people aged 60 and over (International Agency for Research on Cancer, 2020). This pattern appears to be very consistent across different countries (Halso- och Sjukvard, 2012; Ministry of Health, 2016; Cancer Research UK, 2018), and is illustrated in Figure 2.5 for the UK, Sweden and NZ. It can be seen that there are small differences in the incidence rate in the ages 80-<85 and 85-<90.

In NZ literature concerning epidemiology of CRC, there is a common methodological problem: namely, categorisation of age in very broad age groups (10 years or even wider age groups). Such broad age bands are not desirable in CRC research as CRC risk changes with increasing age and the risk in 50 years old is different than in 59 years old. When broad age categories are used, despite the fact that data are available in one-day

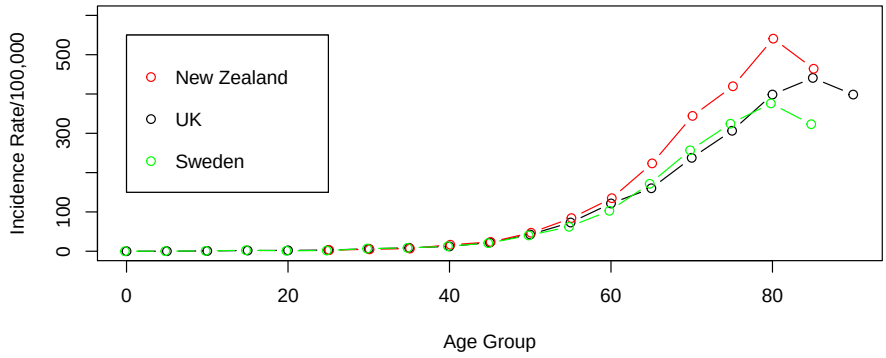


Figure 2.5: *CRC incidence rates as a function of age in New Zealand, UK and Sweden in 5-year age groups which are indicated by the lower bound of each group (Halso- och Sjukvard, 2012; Cancer Research UK, 2018; Ministry of Health, 2016)*

resolution, the data are analysed under the assumption that the risk is constant within the age group. Additionally, incorrect modelling of age can lead to misleading statistical inference, especially when the risk factor under investigation is strongly confounded with age. An example of such risk factor is diabetes status which is strongly correlated with age.

2.5.4.2 Gender

Male gender is a risk factor for CRC. According to Kaminski et al. (2014), male gender increases the risk of CRC at a level similar to a positive family history of CRC. The male-to-female incidence ratios differ between countries (Figure 2.6). In NZ, the male-to-female ratio is closer to one than in most other countries. In the reviewed literature, I did not find any explanation for the reason of the high incidence rates in NZ females vs males, however the variability can reflect exposure to similar risk factors and protective factors in males and females in NZ which might be not true in countries where females have very different lifestyle to males (e.g. Asian countries). The high incidence

rates among New Zealand's females lift the overall incidence in NZ to one of the highest in the world, as the incidence rates in males are comparable to those in other Western countries (International Agency for Research on Cancer, 2020).

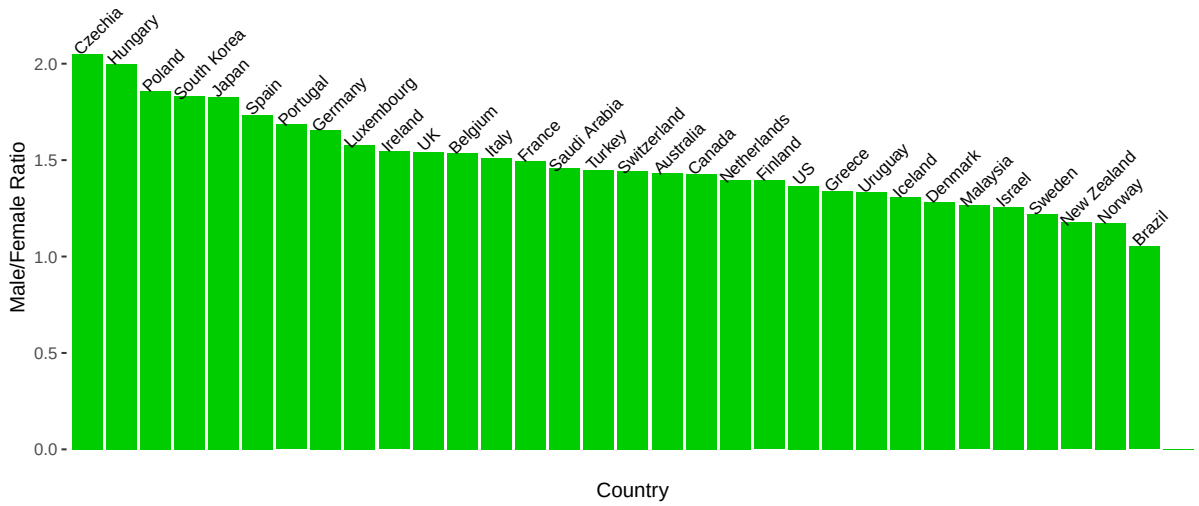


Figure 2.6: *Male to female CRC incidence rate ratios for the 32 countries included in NCI report in International Agency for Research on Cancer (2012).*

2.5.4.3 Ethnicity

Ethnic variation in the epidemiology of CRC within the NZ population has been recognised by many authors (Hill et al., 2010; Sammour et al., 2010; Blakely et al., 2011; Shah et al., 2012; Swart et al., 2013; Moore et al., 2015; Sharples et al., 2018). With respect to incidence rates in Māori and non-Māori, historically, the incidence rates have been reported to be about 50% lower in Māori than in non-Māori (Blakely et al., 2011). The result of an analysis of incidence data from years 2002–2006 by Moore et al. (2015) showed about 35% lower incidence of CRC in Māori vs non-Māori, with standardised rate ratios (SRR) of 0.71 (95% CI; 0.64, 0.78) in males and 0.63 (95% CI; 0.56, 0.70) in females. However, in 2017 the rates differed only slightly, with ASR of 37.3 per 100,000 in Māori

and 39.8 in non-Māori, as the rates have been converging in the last decade (Figure 2.7) (Ministry of Health NZ, 2019d). Although NZ authors claim that the rates in Māori have been increasing at a faster speed than in non-Māori (Blakely et al., 2011; Shah et al., 2012; Teng et al., 2016), it is difficult to understand why studies claim an increasing risk in Māori even after 1994, considering that ASR stratified by gender and ethnicity available from MoH do not show any such trend and, to the best of my knowledge, there is no published study analysing long term incidence trends that could provide such evidence. Finally, as discussed in Section 2.5.2, Māori CRC patients are undercounted in NZCR prior to 2006.

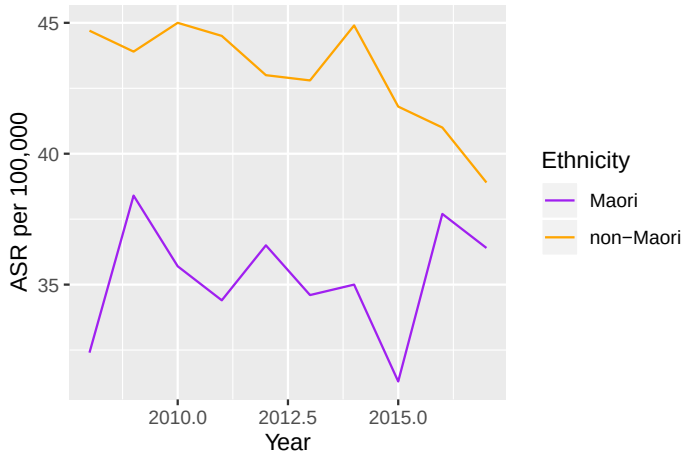


Figure 2.7: *CRC incidence rates in Māori and non-Māori for years 2008-2017 age standardised to the WHO 2000 World Standard population, based on tables from Ministry of Health NZ (2019d).*

Several NZ studies have reported that Māori are younger than non-Māori when diagnosed with CRC (Sammour et al., 2010; Swart et al., 2013). The most recent and comprehensive study investigating barriers to earlier diagnosis of CRC in NZ, the PIPER study, in the report to the Health Research Council included a similar conclusion (Secker et al., 2015). The authors note, however, that this issue still requires further research: "The distribution of age at diagnosis differed by ethnicity, with Māori patients tending to be younger than nMnP [non-Māori non-Pacific] patients. Pacific had a larger proportion

under 60 at diagnosis than either Māori or nMnP. These population groups have different age structures from the nMnP population which needs to be considered, and this will be investigated in on-going analyses". However, in the paper which summarised part of the PIPER study results from 2018 (Sharples et al., 2018), the issue was not resolved. There might be several reasons for the the age discrepancies at CRC diagnosis between Maōri and non-Maōri not being clearly explained in the literature. One reason might be the complexity of research that could explain the phenomenon which I briefly outline below.

While it is true that, in all studies, the reported median age at CRC diagnosis in Māori is much lower than in non-Māori, to test whether the difference in age at diagnosis can be fully explained by the population structure, the hypothesis must be formulated precisely as it is not even clear what researchers mean when they state that Māori are diagnosed with CRC at younger age than non-Māori. One possible meaning of the statement relates to the expected age at diagnosis of CRC for a given birth cohort (longitudinal age curves), assuming that there is no competing risks. The null hypothesis would then be: given individuals from the same birth cohort, otherwise the same but with different ethnicities, the median age at diagnosis would be the same in both ethnicities. To obtain longitudinal age curves the age effect has to be adjusted for cohort and period effects, as the three effects are connected and have to be included in a single model (Rosenberg et al., 2014). This can be achieved by using age-period-cohort analysis of incidence data which is not a trivial task.

With respect to survival in Māori vs non-Māori, despite lower incidence, once Māori are diagnosed with CRC, they die more often as a result of the disease (Jeffreys et al., 2005; Hill et al., 2010; Firth et al., 2016; Sharples et al., 2018). The reason for the disparity is not entirely clear. Sharples et al. (2018) attributed the discrepancy in survival primarily to the characteristics of the tumour (stage, differentiation and location), while Hill et al. (2010) attributed lower survival in colon cancer in Māori to comorbidities, poor post-diagnostic

care and worse access to the health services, rather than to the tumour characteristics at diagnosis.

2.5.4.4 Family history of CRC

Around 30% of CRC patients have a family history of CRC (Butterworth et al., 2006). A meta-analysis by Butterworth et al. (2006) estimated that patients with a first-degree relative affected with CRC suffer around twice higher risk, compared to those with no family history of CRC. Having multiple relatives with CRC increases that risk further.

Genetically determined susceptibility to CRC is not addressed in this study, as patients with genetic susceptibility to CRC are under medical surveillance, and they will not be assessed for presence/absence of CRC using the pathway through primary care and therefore only family history of CRC is covered in this research.

2.5.4.5 Diabetes mellitus

Diabetes mellitus (DM) is a group of chronic metabolic disorders. Based on underlying pathology, the most common forms of DM are: type 1 diabetes (T1D) and type 2 diabetes (T2D). Type 2 diabetes is more common in adult patients and affects approximately 90–95% of patients with DM (American Diabetes Association, 2013). In the recent decades, prevalence of DM has constantly been increasing world-wide. In 2040 the number of patients with DM is predicted to reach 642 million, compared to 415 million in 2015 (Ogurtsova et al., 2017). Following the world-wide trends, the prevalence of DM in NZ is also expected to increase (Coppell et al., 2013). In the remainder of the thesis *diabetes* has the same meaning as *DM*; both are used interchangeably.

The estimated number of people with diabetes in NZ in 2018 was 253,000 (Ministry of Health NZ, 2019e). However, diabetes is an under-diagnosed disease, and accord-

ing to MoH (of Health, 2014) the prevalence could be underestimated by around 40%. Diabetes is associated with male gender and is twice more prevalent in Māori than in non-Māori (Atlantis et al., 2017). The age at diagnosis of diabetes has been declining in the last decade and the trend is expected to continue (Best Practice Advocacy Centre New Zealand, 2018).

CRC is diagnosed in patients with diabetes more often than would be expected by chance only (González et al., 2017), but the link between diabetes and CRC is not fully understood. There are hypotheses for plausible biological links between both diseases; in particular hyperinsulinaemia, possibly hyperglycemia and inflammation which have been backed up with sound evidence (Chang and Ulrich, 2003; Berster and Goke, 2008; Sacerdote and Ricceri, 2018). Additionally, diabetes, in particular T2D, and CRC share common risk factors such as lower abdominal obesity, systemic inflammation, lack of physical activity and diet (González et al., 2017; Gillies et al., 2019) and therefore it is biologically plausible that they often coexist.

Many studies were carried out to investigate the associations between both diseases, often addressing the question whether diabetes has causal effect on the development of CRC, a question that still remains unanswered (Giouleme et al., 2011). Several meta-analyses of observational and case-control studies reported a very similar pooled effect size of about 30% for the increased risk of CRC in patients with diabetes compared to those without diabetes, or to the general population (Larsson et al., 2005; Jiang et al., 2011; Starup-Linde et al., 2013; Wu et al., 2013; Luo et al., 2016; Sacerdote and Ricceri, 2018). Except Larsson et al. (2005), all meta-analyses reported heterogeneity between studies which implies that the association between diabetes and CRC may differ between different populations.

CRC was also reported to be one of four cancers (among 20 cancer sides investigated) for which there was robust evidence for the association with diabetes in the meta-analysis by

Tsilidis et al. (2015). The author stressed the need for inclusion of diabetes duration in association studies which often did not include the term in analyses. The importance of modelling duration of diabetes as time-varying covariate in association studies have been already reported earlier by Carstensen et al. (2012) and Johnson et al. (2012), however many investigators did not include duration of diabetes in their analyses. This serious methodological omission could lead to detection bias and to the overestimation of the strength of some of the reported associations (Carstensen et al., 2012; Johnson et al., 2012).

Most recently conducted cohort studies evaluated the effect of duration of diabetes (Harding et al., 2015; Liu et al., 2015; Peeters et al., 2015; Wang et al., 2015; Valent, 2015; Dankner et al., 2016; Gini et al., 2016; Ballotari et al., 2017; de Kort et al., 2017) and reported consistently increased risk of CRC in persons with diabetes but with varying effect sizes between 1.25 and 1.60.

In addition to duration of diabetes, use of insulin for controlling diabetes has been shown to modify the risk of CRC. Studies which included insulin use in analysis reported higher incidence rates in insulin-users compared to non-users or to users of different medicine for controlling hyperglycaemia (Yin et al., 2014). There is however a controversy about whether insulin has a causal effect on CRC risk (Hernandez-Diaz and Adami, 2010). It was postulated that type of diabetes therapy can also be seen as an indirect marker for clinical severity of diabetes, and the increased risk of CRC in patients using insulin compared to non-insulin users might be a result of confounding by indication bias where patients who use insulin are more ill and, therefore, the association could be increased not as a result of treatment with exogenous insulin only as a result of severity of diabetes and therefore an overall poor health (Yang et al., 2004; Limburg et al., 2005). However, Limburg et al. (2005) used the type of therapy for controlling hyperglycaemia in the analysis as an indirect marker for the clinical severity of diabetes but the study found no association between the marker and incident CRC.

Country-specific factors may play a role in the association between diabetes and CRC (González et al., 2017). These authors investigated the prevalence of diabetes and the incidence of CRC on a worldwide basis and found a lack of correlation between prevalence of diabetes and CRC incidence. Countries in Latin America and the Middle East tend to have high prevalence of diabetes but low incidence of CRC, while in African countries, prevalence of diabetes and CRC incidence are both low. In Western countries, including NZ, where the incidence of CRC is high, diabetes prevalence is intermittent. However, among Western countries with high CRC incidence, NZ has one of the highest reported prevalences of diabetes.

Interestingly, González et al. (2017) concluded that if the association between diabetes and CRC reported in many studies holds, the epidemic of diabetes fuelled by changes in lifestyle may trigger a wave of CRC diagnoses. However, the correctness of the conclusion could be questioned. Despite the well evidenced positive association between diabetes and CRC, the increase in prevalence of diabetes is actually more likely to reduce CRC incidence. This is because patients with diabetes have a shorter lifespan than individuals without diabetes as shown in Carstensen et al. (2014) who studied the lifetime risk of CRC in the Danish population and reported the lifetime cumulative risk to be lower in patients with diabetes than in patients without diabetes. Because CRC risk is higher in old age, if more people die at a younger age there will be fewer individuals in the population who could be diagnosed with CRC. On the other hand, the hypothesis presented by the authors could be correct as patients with diabetes have been getting better medical care resulting in better control of hyperglycaemia which, as a consequence, could increase lifespan in patients affected by diabetes. If so, the incidence of CRC due to the increasing diabetes prevalence might indeed, as concluded by González et al. (2017), rise in the future.

In summary, the positive association between diabetes and CRC seems to be universally agreed but the strength of the association differs between populations. My searches of literature did not identify any study which investigated the association in the NZ popula-

tion, despite the fact that diabetes is so prevalent in NZ. Considering that diabetes status has already been proposed as an additional factor that can be used in clinical practice to assist doctors during the investigation of patients for presence/absence of CRC (Giouleme et al., 2011; Starup-Linde et al., 2013), the knowledge about the association in the NZ population would clarify whether NZ physicians should also be particularly attentive to patients with diabetes when investigating patients for presence/absence of CRC. Estimating the association is especially important for the Māori population, given the twice higher prevalence of diabetes compared to non-Māori (Atlantis et al., 2017).

2.6 Ways to reduce the CRC burden

The previous section shows the magnitude of the CRC burden in NZ compared to the other developed countries and highlights differences in the incidence rates and survival between Māori and non-Māori. It is accompanied by a lack of research related to CRC incidence that could explain the differences between the two main NZ ethnicities and thereby help in tackling the ethnic inequalities in CRC outcomes. Also, there has been indicated a fast growing CRC incidence in young New Zealanders in recent years but the research on that topic is also scarce. Additionally, NZ faces a problem with limited resources for colonoscopy. However research that could help in the prioritisation of patients for colonoscopy and could assist physicians with selection for investigation those who are at highest risk is also scarce.

Addressing the CRC burden is a task which combines the efforts of government bodies, the medical sector and scientists. In the next section I briefly present some of the undertaken initiatives and also explain the role of cancer surveillance research in the fight against CRC.

2.6.1 Initiatives for tackling CRC

CRC is one of over 200 types of malignant diseases which are a cause of premature deaths and a burden to the whole world population. To lower the burden of cancer, initiatives such as the Cancer Moonshot in the US (Lowy and Singer, 2017) and the initiatives undertaken by the European CanCer Organisation (ECCO) (Lawler et al., 2019) were developed. As the European health-care systems are different from the US system, the ECCO committee decided to adapt the US model according to its own principles and needs. Similarly, any initiative in NZ for tackling the burden of cancer would not necessarily fully mimic the aims of initiatives from other continents. Instead, it should be grounded in NZ realities and based, where possible, on analysis of data collected from the NZ population. The main goals of such initiatives, however, are universal, and include accelerating the understanding of cancer and its prevention, early detection, treatment, and cure. An increased access to new research, data, and computational capabilities are necessary tools to achieve these goals and they will be similar in different countries (Lowy and Singer, 2017). So for example, the appropriate tools for analysis of incidence data from most cancer registries will be the same.

When it comes to tackling cancer in NZ, in 2003, the government published the *New Zealand Cancer Control Strategy* document which described the need for reducing the cancer burden, stressing the need for research in NZ (Ministry of Health and the NZ Cancer Control Trust, 2003). The document states that, apart from implementation of population-based screening for CRC, there is also a need for less resource intensive strategies to reduce the delay in CRC diagnosis and facilitate diagnosis in the early stages to achieve better survival. The document stresses the need for research and improvement without putting a stress on already limited public resources.

Although NZ does not have a primary care electronic database which could help to provide an evidence base for the Ministry of Health's policy, the MoH holds good quality registry

data. This data can be analysed to answer many research questions (Parkin, 2008) that can help in setting up policies which will be beneficial to the whole population. Cancer surveillance research, explained in more detail in the next section, is designed to use such data to provide valuable and powerful results (Glaser et al., 2005). In a recent NZ paper, Sarfati and Jackson (2020) commented on the lack of progress in cancer control strategies in NZ and noted missed opportunities in prevention since the *New Zealand Cancer Control Strategy* document was released in 2003. In the case of CRC, the lack of relevant NZ epidemiological studies could be one of the reasons for the lack of the desired progress.

2.6.2 Cancer surveillance research

Cancer surveillance research is a discipline in epidemiology which analyses and interprets systematically collected data included in cancer registries and population data, in order to generate and/or test hypotheses about cancer incidence, cancer outcomes and cancer predictors in well-defined populations over time (Glaser et al., 2005). Cancer surveillance research has been acknowledged as having great potential for helping to decrease the cancer burden. This is achieved through contributions to the reduction of cancer incidence (by identifying risk factors and testing hypotheses about causes of cancer), and to improvement of patient survival after diagnosis (by improving earlier detection of cancer). Age-period-cohort models are well accepted tools for analysis of data in this field (Smith et al., 2016; Rosenberg, 2019). I provide a detailed explanation of an APC model in Chapter 3 (Section 3.1.2).

By analysing data from cancer registries and population data, cancer surveillance research can address a wide range of research questions and provide results that are generalisable to the whole studied population as the analyses are conducted using population level data (Glaser et al., 2005). In the past, the analyses of trends using data from cancer

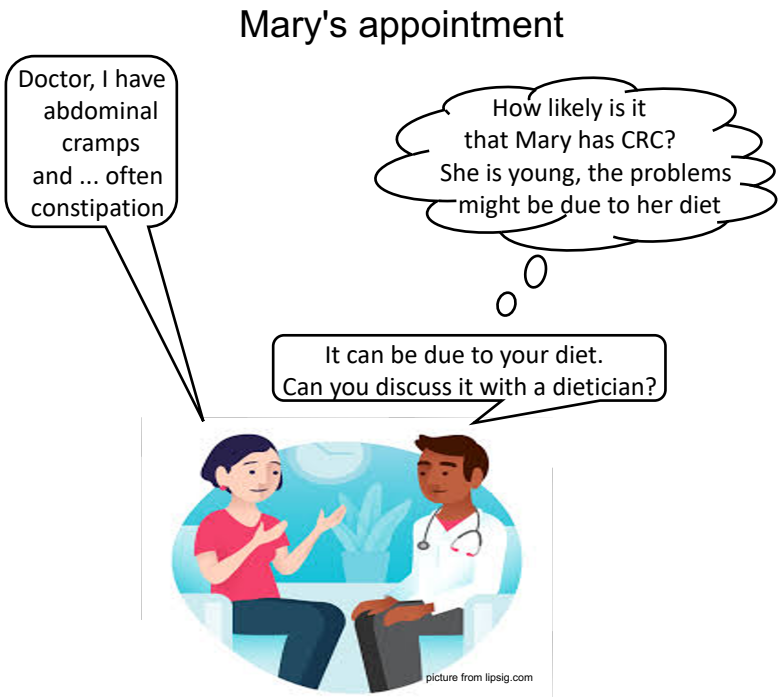
registries using APC models led to important findings, such as the carcinogenic properties of the drug diethylstilbestrol which, when used in mothers during pregnancy, was found to be associated with vaginal cancer in their daughters (Hoover et al., 2011). Another example is research conducted in Norway by Svensson et al. (2002) who found that CRC incidence decreased in generations that were exposed to calorie restrictions during World War II.

The results from analysis of long-term trends can help in making projections for the future incidence of cancers (Bray and Møller, 2006). Such projections can be used e.g. in health economics to estimate the cost related to cancer and for management of future resources as presented in Sheerin et al. (2015). This NZ study made projections for 2014-2026 assuming that the growing incidence trend in Māori will reverse, and the downhill linear trend in non-Māori will persist based on incidence data from the previous 5-year period only. Such assumptions would be more reliable if the analyses of trends were based on much longer period.

In NZ, since 1994, the NZCR provides good quality nation-wide data about registrations of CRC. Additionally, the MoH holds population-based data which can be accessed by researchers, including VDR for registrations of patients with diabetes, NMDS or PC (explained in Section 1.4.2). Stats NZ provides very well-structured tables with population counts, stratified by gender, ethnicity and data which are not publicly available can be ordered. All this taken under consideration, NZ has great potential for conducting cancer surveillance research, however the resources have until now been under-utilised in the case of the research on CRC epidemiology.

2.7 Summary

Chapter 2 provided background knowledge and literature overview that led me to finding gaps in research which are important to address, as the new knowledge might help to improve early diagnosis of CRC in NZ patients. Importantly, NZ has good-quality population-based and cancer registry data that offer this great potential for researchers. Of particular importance for this study, analysis of registry data can lead to identification of specific strata with increased incidence of CRC. Here, I would like to bring back the fictional patient Mary and again to ask the question: can we provide results based on analysis of already existing population-based and administrative data that could help doctors in earlier diagnosis of patients with already existing but not yet diagnosed CRC?



Can we help the GP to decide how likely it is that she has CRC?

The following three chapters present the empirical part of this study that I conducted in order to answer the question. The empirical part consists of three independent sub-studies (sub-study 1, 2 and 3) that addressed the different objectives stated in Section 1.3. For each sub-study in the following three chapters I provide background and methods which are relevant only for the specific sub-study. Subsequently, the results of analyses are followed by their discussion.

Chapter 3

Age-period-cohort analysis of colorectal cancer incidence in New Zealand for the period 1994–2018

3.1 Introduction

Chapter 3 describes a study of long-term trends in colorectal cancer (CRC) incidence in NZ for the period 1994–2018, using an age-period-cohort (APC) analysis of data from the New Zealand Cancer Registry (NZCR) [1]. The first part of this section provides a basic overview of the trends in CRC incidence in NZ and in selected comparable countries. I then explain the basis of age-period-cohort (APC) analysis, including the most important issues with parametrisation and interpretation of the results of APC analysis. Finally, I discuss the rationale for the sub-study and the objectives it addresses.

3.1.1 Trends in CRC incidence

Analysis of trends in CRC incidence rates over a long time period can improve our understanding of CRC epidemiology. Such analysis can help us to make more accurate predictions of the disease burden, and to identify individuals and population segments that are at high risk of CRC (Bray and Møller, 2006; Rosenberg and Anderson, 2011). The identification of certain strata at high risk of CRC is valuable to both physicians and policy-makers. For physicians, such analysis can inform decisions about prioritisation for further investigation of patients suspected of CRC. For policy-makers, such analysis can contribute to more accurate prognosis of future CRC incidence rates, which can in turn improve the allocation of future resources for diagnosis of CRC and treatment of patients diagnosed with CRC. Trend analysis of the observed incidence rates has already provided valuable information for the assessment of a diagnostic pathway in other countries. In Canada, Singh et al. (2008) investigated trends in the incidence rates of proximal and distal CRC. Their analysis led to two conclusions. First, that flexible sigmoidoscopy (FS)

[1] Parts of this sub-study were presented at the virtual conference of the International Society for Clinical Biostatistics in Krakow in 2020, as an oral poster presentation, and at the Postgraduate Statistical Conference at the University of Waikato, also in 2020.

and faecal occult blood test (FOBT) might be not the best screening methods for those 70 years and older; for these patients a colonoscopy might be a better tool for investigation of CRC, especially in females. Second, that FS and FOBT could have a high diagnostic value in patients younger than 50 years.

Overall trends for age-standardised CRC rates are presented in World Health Organization (WHO) statistics publications every few years (International Agency for Research on Cancer, 2020). Such statistics, based on new CRC registrations per age group and per calendar year, are also produced in many countries, including NZ. Based on tables published yearly by the NZ MoH, the overall incidence rates in CRC have decreased during the last two decades, with similar trends observed in males and females (Ministry of Health NZ, 2018a).

However, the NZ-based study by Gandhi et al. (2017) reported that the trends are different in different age groups. In other words, there are interactions between the age of diagnosis and the year (period) of diagnosis. In particular, Gandhi et al. found an increase in incidence of rectal cancers in young individuals in NZ in recent years. Associations between age and trends in incidence rates have also been reported in other highly developed countries (Brenner et al., 2017; Feletto et al., 2019; Rosenberg, 2019; Araghi et al., 2019). Rosenberg (2019) reported that the rising incidence rates in rectal cancer in patients younger than 50 years in the US population could best be explained as an increasing incidence in specific (late) birth cohorts rather than in specific age groups. Likewise, Brenner et al. (2017) and Feletto et al. (2019) reported cohort effects responsible for the growing CRC incidence in the younger segments of the Canadian and Australian populations, respectively.

Araghi et al. (2019) identified a cohort effect responsible for the increasing incidence in young adults from NZ. Araghi et al. (2019) analysed CRC incidence data from seven high-income countries including data for NZ for the years 1995–2014. For NZ, Araghi

et al. (2019) reported highly increased incidence in colon and rectal cancers in generations born in the 1990s compared to the incidence in those born in 1925. As reported by Araghi and colleagues, the trends in NZ were similar to the trends observed in the UK, Norway, Australia and Canada.

The identification of a cohort effect was possible due to the analysis of data using an age-period-cohort model. APC models can disentangle age, period and cohort effects contributions to the cancer incidence rates, and can investigate whether the observed differences in incidence rates between different age groups are due to age-period interactions, or can be better explained by the cohort effect (Bell and Jones, 2013). The identification of a cohort effect, as opposed to age-period interactions, has major consequences for clinical practice and for public health. This will be discussed in the relevant sections in more detail. Here, it is important to stress the major benefit of using an APC model for analysis of cancer incidence data; the APC model is a simple additive model that (assuming a good fit to the data is provided) can estimate the incidence of CRC for any population strata precisely (Rosenberg and Anderson, 2011). In this respect, an APC model differs from a statistical model with age-period interactions, which would be very complex and could easily lead to overfitting of the data. As cancer registries follow multiple cohorts over a long time, analysis of such data allows researchers to uncover fundamental changes in incidence rates that are difficult to identify in studies using cohort or case-control designs (Rosenberg and Anderson, 2011). Additionally, APC analysis can give clues about etiological factors, which can have a wide and important use in public health for prevention purposes.

APC analyses of CRC incidence data, including data from recent decades, have been carried out in many countries to determine how the trends in CRC incidence can be explained. However, in NZ, where the burden of CRC is so high, such analysis has not been carried out despite the high quality data available from NZCR and Statistics NZ. Age-period-cohort modelling is used by the MoH for making predictions of cancer incidence

rates for the near future; however, I have not found any publication that gives estimates of cohort and period effects and their interpretation in publicly available sources. The MoH uses the APC model as one of four models from which forecasts for the next three to five years are made, which are subsequently averaged to give the final forecast (Ministry of Health NZ, 2010).

Despite the scarcity of published results of APC modelling using CRC incidence data from the NZCR, two papers investigating trends in CRC incidence have recently been published (Shah et al., 2012; Gandhi et al., 2017). However, neither author used APC analysis in their research. Both studies reported a heterogenous picture, with upward trends in some demographic groups and anatomic sub-sites and downward trends in others. The data were tabulated by age and period. If in an age-period model interactions between age and period are included, and the interaction exists (as reported in Gandhi et al. (2017)), this could be due to cohort effect, which unfortunately was not investigated. Analysis which does not include cohort effect can lead to misleading results, attributing an actual cohort effect to interactions between period and age effects. For example, the increased risk with time in patients younger than 50 years reported by Gandhi et al. (2017) might in fact be an increased risk in the relevant birth cohorts. In this scenario, it might appear as if the change happens in young patients only, because for the recent birth cohorts (generations born in 1970s and 80s) the available data were from young age groups only.

While Gandhi et al. (2017) presented interesting findings, it is worth noting that its analysis was based on very crude time scale categories. Gandhi et al. compared trends in three age groups, <50, 50-<80 and 80+, for distal, proximal and rectal cancers in a study period divided into two decades, with interactions between all predictors. Such crude categorisation of age and period was necessary to avoid overfitting the data, due to the high number of interactions included in the model. However, the crude categorisation of age and time scales in CRC research can be problematic, as it may not capture the true relation between the age effect, the period effects, and the CRC incidence.

3.1.2 Introduction to age-period-cohort analysis

The age-period-cohort model is a fundamental model which can help us to understand time-varying elements in data related to cancer incidence and mortality (Smith et al., 2016). The application of APC models in epidemiology began in the 1980s (Holford, 1983; Clayton and Schifflers, 1987a,b), and APC models have since been used continuously as a standard tool in cancer epidemiology (Robertson and Boyle, 1998; Holford, 1991; Dubrow et al., 1993; Carstensen, 2007; Rutherford et al., 2010; Smith et al., 2016; Chernyavskiy et al., 2018; Murphy and Yang, 2018; Feletto et al., 2019; Holford et al., 2019; Rosenberg, 2019). APC models investigate three types of time-varying phenomena that contribute to the incidence rates: age (a) effects, cohort (c) (the year of birth) effects, and period (p) (year of diagnosis) with each time-scale capturing distinct processes (Keyes et al., 2010; Yang and Land, 2013).

When for example the outcome of interest is CRC incidence, age captures processes related to the damage of the colon due to genetic mutations accumulated with increasing age. Alternatively, age can be a surrogate of the accumulated exposure to external risk factors such as smoking or unhealthy diet. The cohort component of the temporal trend can suggest the exposure to factors which were characteristic for the particular generation, often during their early years of life (Keyes et al., 2010; Murphy and Yang, 2018). A birth cohort effect can also be a surrogate for exposure to certain carcinogens even before birth, as in Hoover et al. (2011), who explained how the identification of a cohort of young females with a high risk of vaginal cancer helped to establish a causal link between the use of a specific drug (a synthetic estrogen diethylstilbestol) by females during pregnancy and the risk of vaginal cancer in their daughters. Cohort effect affects only specific generations (not the entire age groups) and is carried out throughout the whole life.

Unlike age and cohort time scales, which are related to biological and environmental

factors causing changes in the body, the period effect is related to the processes by which changes in the body are identified leading to a diagnosis of CRC (Yang and Land, 2013; Smith et al., 2016). A period effect can reflect for example changes in diagnostic procedures or disease classification, and affects all age groups simultaneously.

Studies using APC models have improved our understanding of the burden and aetiology of several cancers including colorectal cancer (Murphy and Yang, 2018). An example is a study by Svensson et al. (2005), which identified lower CRC incidence rates in generations born during World War II in countries where calorie restriction was implemented. Identification of a cohort effect also helped to establish risk factors in different cancers, e.g., for lung cancer, younger age at smoking initiation and longer duration of smoking (Jemal et al., 2003; Bray and Weiderpass, 2010), and also in other diseases e.g., identification of the effects of age and birth cohort on the increasing rates of obesity in Australian females (Dobson et al., 2020).

APC models are fitted to estimate age-specific rates in disease incidence, adjusted for cohort and period effects, and to estimate the cohort and the period effects themselves. Additionally, APC models can be used to estimate incidence rates in specific population strata (Rutherford et al., 2010). Such estimates are relevant for many decisions in health care, e.g. assessing the cost effectiveness of screening programs or diagnostic tests (Salzmann et al., 1997).

To fit an APC model for incidence rates of a studied disease, a tool called Lexis diagram is used. The next sections introduce Lexis diagram and then discuss the mathematical basis of an APC model.

3.1.2.1 Lexis diagram for incidence rates

Incidence rates are calculated by tabulation of cancer incidence data (often obtained from cancer registries) and population counts (mostly obtained from censuses). A table which combines cancer incidence counts and the population values can be summarised in a Lexis diagram (Figure 3.1) (Keiding, 1990). Each cell of the Lexis diagram corresponds to the combination of a calendar period, a birth cohort and an age group, and contains cancer incident count and population count expressed in person-years.

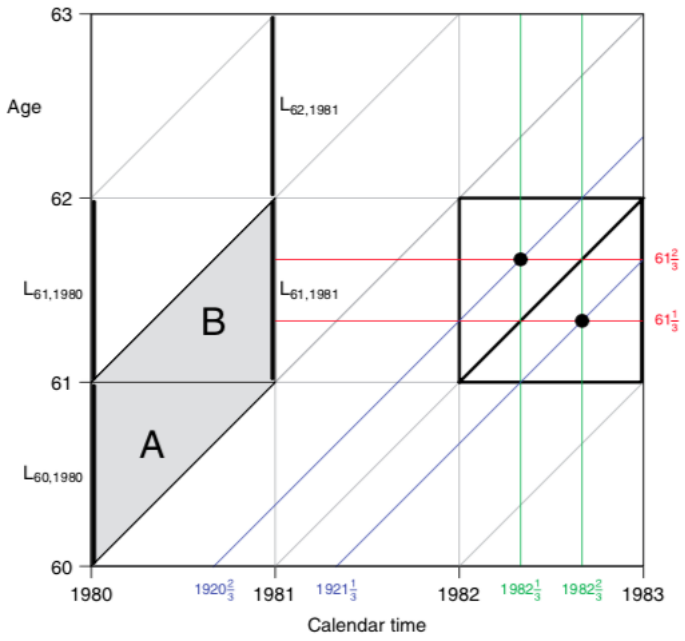


Figure 3.1: *Lexis diagram reproduced from Carstensen (2007). The thick lines in the left part show the population figures at the beginning of 1980 and 1981 necessary to estimate the population risk time in the triangles A and B. The right part of the diagram shows the mean age, period and cohort in the triangular subsets of a Lexis diagram. Note the connection between age, period and cohort: p=c+a: 1982 =1920.333 +61.666 and 1982.666 =1921.333 +61.333 (Carstensen, 2007).*

In cancer epidemiology, data tabulated by age and period are used in: standard non-parametric descriptive and exploratory methods, such as graphical presentation of cancer

rates, tables with age-standardised cancer rates or joinpoint regression; and parametric statistical models i.e., age-period-cohort models, that can test and also generate hypotheses (Keiding, 1990; Last et al., 2001; Carstensen, 2007; Rosenberg and Anderson, 2011). An example of using results from an APC model to generate a hypothesis would be, first, the identification of particular birth cohorts with increased incidence rates, and, second, the generation of a hypothesis about risk factors that affected those particular birth cohorts. Hypotheses generated in this way can subsequently be explored in future studies to give clues about the aetiology of CRC (Rosenberg and Anderson, 2011).

3.1.2.2 Mathematical basis of the APC model

The mathematical foundation of APC models is explained in Holford (1983). This outlines the basics of the methodology used in this sub-study. The additive APC model is based on the assumption that for a given observational unit (Lexis cell) the log rate is a sum of age, cohort and period effects. Equation 3.1 shows the general form of the APC model for rates λ $(a,\ p)$ at age a in a period p, in birth cohort c=p–a.

$$log[\lambda_i(a_i, p_i)] = f(a_i) + g(p_i) + h(c_i) \tag{3.1}$$

where: a_i, p_i and c_i represent the mean value of each time-scale for the Lexis cell i

The model allows the effects of age, period and cohort to be non-linear, and therefore each effect can be broken down into a linear and a non-linear component:

$$f(a) = \tilde{f}(a) + \mu_a + \delta_a a \tag{3.2}$$

$$g(p) = \tilde{g}(p) + \mu_p + \delta_p p \tag{3.3}$$

$$h(c) = \tilde{h}(c) + \mu_c + \delta_c c \tag{3.4}$$

where: $\tilde{f}(a)$, $\tilde{g}(p)$, $\tilde{h}(c)$ are the non-linear components called deviation or curvature, μ_a, μ_p, μ_c are the intercepts, and δ_a, δ_p, δ_c are the slopes.

However, due to the linear dependency of the three time scales, the model is over-parameterised leading to an identifiability problem discussed in the next section.

3.1.2.3 The identifiability problem in APC models

There is a fundamental problem of APC analysis, namely the well known lack of identifiability due to the exact linear dependency of the three time scales; that is, knowing two of the three terms, the third can be calculated from the equation period = cohort + age (Holford, 1991; Robertson and Boyle, 1998). It is therefore not possible to estimate all nine components of the three time-scales (equations 3.2, 3.3 and 3.4) at the same time because there is an infinity of solutions for this model. In order to make the model identifiable, it is necessary to impose some constraint on the model. In cancer epidemiology, it is common to set functions g and h to zero for a chosen reference cohort and reference period and to set the period slope to zero. Setting the period slope to zero is a simple and biologically plausible solution used in earlier research (Larsen and Bray, 2010; Araghi et al., 2019). Since there is no mathematical rationale for the choice of particular constrains, the choice has to be based on the understanding of the factors which play a role in diagnostic procedures and the disease itself (Dubrow et al., 1993; Bell and Jones, 2013).

The explanation of the rationale for setting the period slope to zero can be seen in the following example: a period effect could be caused by e.g. increased incidence rates of CRC due to initiation of a population-based screening. Initially, incidence rates will increase in age groups eligible for screening. However, in the following years, as the pool of existing but not yet diagnosed CRCs will be depleted, the initial increase will most likely be followed by a drop in incidence rates. Hence, over a longer time, the period effect will likely have a negligible trend. In this case, the solution of setting the period slope to zero would, therefore, be reasonable.

Depending on the parametrisation used, the estimates for the age, cohort and period effects will be different. The choice of parametrisation, therefore, has implications for the interpretation of the results (Carstensen, 2007; Smith et al., 2016). However, the predicted incidence rates are not affected by the parametrisation, which is an important feature of the APC model, especially when predictions are intended to be used in clinical practice (Carstensen, 2007). There are several different approaches to APC modelling, but all methods have the identifiability problem (Bell and Jones, 2013).

In addition to age, period, and cohort effects which depend on the parametrisation, APC models provide functions that do not depend on parametrisation: the cohort deviations and the period deviations (curvature or non-linear components), net drift and local drifts. Those functions can be explicitly identified and are, therefore, referred to as estimable functions (Rosenberg and Anderson, 2011).

3.1.2.4 Other issues with APC models

In addition to the identifiability problem, APC analysis can have a number of other issues. The classical APC analysis fits the effects of age, period, and cohort as factors, but this approach does not acknowledge the correlation between the cancer rates in adjacent age and period groups (Rutherford et al., 2010). Additionally, in the classical approach, even if data are available in 1-year age brackets it is common practice to reduce the data to 5-year brackets to avoid large fluctuations in cancer rates in adjacent years. This practise results in overlapping cohorts (i.e. if a person is diagnosed in the 2005–2009 period interval at the age of 50–54 years, the person's birth cohort might be as early as 1950 and as late as 1959), which can cause a lot of noise in tabulated data. These issues can be solved, fully or partially, by the use of advanced statistical methodology.

For example, the age, period and cohort effects can be modelled using parametric and non-parametric smooth functions to address the correlation between adjacent years. Further,

the overlap between cohorts can be avoided by adding birth year as a third dimension to the tabulation in the Lexis diagram which divides each square into two triangles (Durrleman and Simon, 1989; Carstensen, 2007). New papers are still debating different solutions to the problem of noise in data in APC modelling. For example, Chernyavskiy et al. (2018) discusses correlation of rates in the neighbour Lexis cells, while Rosenberg (2019) uses smoothing functions as a superior method to categorical modelling of the three time-scales. To the best of my understanding, Carstensen (2007) proposed a method with solutions to the above mentioned problems, and implemented it in the *Epi* package (Carstensen et al., 2019). An elaborate review and critique of APC modelling can be found in Fosse and Winship (2019).

3.1.2.5 Modelling age, period and cohort effects

Using APC models, each of the three timescales (age, period and cohort) can be modelled in many different ways, e.g. as factors (the classical APC model); (Rosenberg et al., 2014), as linear functions using power transformation (Rosenberg, 2019); or as splines (Carstensen, 2007; Larsen and Bray, 2010; Rutherford et al., 2010; Dobson et al., 2020). In the literature which I reviewed, the most common solution to the modelling of time-scale effects is the classical approach.

However, Carstensen (2007) argued that modelling age, period and cohort effects should be carried out using parametric smooth functions, due to the continuous nature of the original scales. Different types of splines, natural splines or fractional polynomials can be used, and any of those approaches will give similar results if sufficient data are available (Carstensen, 2007). In approaches which use splines to model time-scales, the choice of the numbers of parameters for modelling each of the time-scales (number of knots) and the location of knots is an important consideration, as it will have an effect on the estimated incidence rates. Unlike Heuer (1997), who suggested using one knot per five

years, in Carstensen's view the number of CRC cases between knots should be considered when deciding on the location of knots. Carstensen suggests including the same number of events between knots (e.g. quantiles) instead of using fixed numbers of years between knots, so that the number of cancer cases are distributed equally between knots.

With respect to the number of knots for each time-scale, Carstensen suggested using a similar number of knots for all three timescales; however, different numbers of knots for each time-scale can be used, if appropriate. For example, if the period time-scale is expected to show a steep rise, it can be modelled using more knots to allow the model to show the trend in the data. It is desirable to allow the model to be complex enough to show the nonlinear relationship between a particular time-scale and the incidence rates, but without allowing a too wobbly curve in order to prevent overfitting the data. With respect to the location of knots, Carstensen (2007) suggests including the same number of events between knots (e.g. quantiles) instead of using fixed numbers of years between knots. This would distribute the number of cancer cases equally between knots, due to the proportionality between the amount of information in the data and the number of events. However, according to Heuer (1997) knots should be placed equidistantly. Although both authors discuss the issues of selecting the number of knots, neither propose a rule for that aspect of the model specification.

3.1.2.6 Interpretation of the results from APC models

The output of an APC model provides parameter estimates that describe the relationships between the (incidence) rate of a disease and the age of the studied population, the period of diagnosis, and the year of birth. The interpretation of the results given by APC models is complicated, as the modelling requires 'strong assumptions that may not hold in practice' (Holford et al., 2019). Smith et al. (2016) provide an overview of common issues with the interpretation of the results from APC models, concluding that several

authors misinterpreted the output given by APC models. In order to understand better the output from the APC model prior to analysis of cancer rates using APC models, it is strongly recommended that researchers conduct a visual analysis of descriptive graphics of the trends, before the constrains are imposed in model (Holford, 1991; Carstensen, 2007; Smith et al., 2016; Holford et al., 2019).

3.1.3 Rationale

Research investigating long-term trends in CRC incidence in NZ data is scarce, especially with respect to age-period-cohort analysis. In particular, age, cohort and period effects by ethnicity have not previously been investigated. Studies which described long-term trends in CRC incidence used very crude categorisation of period and age scales, despite the availability of good quality data in one-year resolution. By disentangling age, cohort and period effects in incidence data, our knowledge about factors affecting incidence can contribute to addressing the burden of CRC in NZ.

3.1.4 Aim and objectives

Considering the lack of research investigating temporal patterns in CRC incidence rates in NZ using high-resolution time scales, the sub-study aimed to fill this gap by addressing the following objectives:

1. Describing the overall patterns and trends in CRC incidence rates in NZ from 1994–2018.

2. Fitting an age-period-cohort model to disentangle the contribution of age, period and cohort effects to the temporal trends of CRC.

3. Using age-period-cohort modelling to estimate age-specific incidence rates in the

population strata for use in clinical practice.

To address these aims, I conducted a population-based study using CRC incidence data from the NZCR.

3.2 Methods

3.2.1 Data and study design

This was a population-based observational study which analysed incidence CRC data from the NZ population. This sub-study included all incident CRC patients 30-<90 years old, registered between 1994 and 2018 in the NZCR with ICD codes C18.0–C 20.0, with a registration in NZCR used as diagnosis marker. Cancers located in the cecum (C18.0), appendix (C18.1), ascending colon (C18.2), hepatic flexure (C18.3) and transverse colon (C18.4) were categorised as proximal, splenic flexure (C18.5), descending colon (C18.6) and sigmoid colon (C18.7) were categorised as distal, and recto-sigmoid junction (C19) and the rectum (C20) were classified as rectal. Patients with only overlapping (C18.8) and/or tumours with unspecified location (C18.9) were classified as unspecified location, and they were not included in sub-site specific analyses. Diagnosis years 1994–2018 were used due to data availability for the whole PhD project. Additionally, in 1994 registration of all cancers became mandatory due to the introduction of the Cancer Registry Act (Ministry of Health NZ, 2020b), and therefore from 1994 the registration of CRC in the NZCR can be considered complete.

The corresponding counts of the entire NZ population in 1-year age brackets at 31 December of each year, stratified by gender and ethnicity, were sourced from Statistics NZ (Statistics NZ, 2020a).

3.2.1.1 Calculation of person-years

Calculation of person-years was carried out using the *N2Y* function from the *Epi* package (Carstensen et al., 2019). Tables with count of the population were used to estimate the number of person-years in each age-period-cohort combination using the following formula for the Lexis triangle A in Figure 3.1:

$$y_{a-1,p,p-a} = \frac{1}{3}L_{a-1,p} + \frac{1}{6}L_{a,p+1} \tag{3.5}$$

and for the Lexis triangle B in Figure 3.1:

$$y_{a,p,p-a} = \frac{1}{6}L_{a-1,p} + \frac{1}{3}L_{a,p+1} \tag{3.6}$$

where: L is the population size in age a at the beginning of year p,
$y_{a,p,p-a}$ is the risk time in age a during year p for cohort p-a

These formulas are based on the assumption of a constant mortality in each triangle of the Lexis diagram (Carstensen, 2007).

Rosenbauer and Strassburger (2008) noted that the rates for Lexis triangle using formulas 2 and 3 are incorrect and suggested the following improved formulas:
for the Lexis triangle A

$$y_{a,p,p-a+1} = \frac{3}{8}L_{a,p} + \frac{1}{8}L_{a+1,p+1} \tag{3.7}$$

and for the Lexis triangle B

$$y_{a+1,p,p-a-1} = \frac{1}{8}L_{a,p} + \frac{3}{8}L_{a+1,p+1} \tag{3.8}$$

Although Rosenbauer and Strassburger (2008) expected only small discrepancies between rates calculated using formulas 3.5 and 3.6, and rates based on their formulas, to determine how sensitive the APC model was to this issue I conducted an additional analysis in

which the formulas 3.5 and 3.6 were replaced with formulas 3.7 and 3.8. The discrepancies in my data were negligible, and therefore the calculation of person-years was carried out using the original version of the *Epi* package.

The data tabulated in 1-year intervals for age and period according to the birth cohort (leading to the triangular structure of the Lexis diagram) gave a total of 3000 Lexis triangles [(90-30)x(2018-1994)=1500 different age-period categories (Lexis squares), each of which was further subdivided by date of birth into two categories (Lexis triangles)]. For additional analyses of CRC incident trends in population strata, separate tables for males, females, Māori and non-Māori were constructed - each of which contained 3000 Lexis triangles with population risk-time and the number of CRC cases. The stratified tables were used for fitting separate models for each gender and ethnic group.

3.2.1.2 Tabulation of CRC cases

All CRC cases were tabulated into Lexis triangles based on age at diagnosis, year of diagnosis and year of birth. For the sub-site specific analyses, separate Lexis tables were constructed for proximal, distal and rectal cancers for each gender. Separate tables for Māori and non-Māori were also constructed. Due to the undercounting of Māori ethnicity in NZCR before year 2006 (Shaw et al., 2009), for the purpose of sensitivity analysis carried out to assess the impact of the undercounting on the results, additional tables with corrected counts of Māori and non-Māori were also constructed by multiplying the uncorrected counts in the Lexis cells for Māori by the following correction factors: for years 1994–1995 the factor was 1.31; for years 1997–1999 the factor was 1.48; and for 2001–2004 the factor was 1.31 (Shaw et al., 2009). For years 2006–2011 I did not apply any correction due to the small value of the correction factor (equal to 1.01) (Boyd et al., 2016). Applying such small correction factor to the Lexis cells for Māori would have no effect as the highest count for a single cell is 9 CRC cases and after rounding

the correction would have no effect. For years 1996, 2000 and 2005 I used the midpoint value between the adjacent years (1.395, 1.395 and 1.16 respectively). Those factors were given by Shaw et al. (2009) specifically for CRC (the correction factors differ by cancer type). After year 2011 I did not apply any correction factor as according to Boyd et al. (2016) there was only a minimal disagreement between NZCR and census data after this period. Subsequently, Lexis cells with corrected counts of CRC cases for non-Māori were calculated by subtracting the corrected counts for Māori CRC cases from the total CRC counts.

3.2.1.3 Quality check of data preprocessing

To check if data preprocessing was performed correctly, the following quality checks were carried out: the number of person-years and CRC cases in ten randomly selected Lexis triangles were compared to the source data, and the total number of CRC cases in each stratum defined by gender, ethnicity or anatomical location was calculated from Lexis triangles and compared to the source data.

3.2.2 Statistical analysis

The statistical analysis was conducted using the method described by Carstensen (2007), who included a detailed explanation of the implementation of APC modelling, including explorative data analysis and presentation of results, as implemented in the *Epi* package for R (Carstensen et al., 2019).

3.2.2.1 Descriptive statistics

Characteristics of all CRC patients, and separately for patients with tumours located in the proximal colon, the distal colon and the rectum, were given by gender, ethnicity,

Duke's stage at diagnosis, and by the following age groups: 30-<60, 60-<75 and 75-<90, which relate to the age for population screening in NZ. Continuous variables were summarised as median and IQR, and discrete data as count and percentages. Stage at diagnosis is not used in the statistical analysis; it is given only for descriptive statistics, and presented for periods from 1999 to 2018, as in the previous years different staging classification was used in NZCR, and most likely the stages are not comparable.

To explore the overall time trends in incidence rates before proceeding to model-based analyses, age-standardised incidence rates for each year from 1994 to 2018 were plotted. The age-truncated incidence rates were calculated using direct standardisation to the age structure of the 2018 NZ population aged 30-<90 years. This standard population was chosen because 2018 is the most recent census year. Using a NZ standard population prevents comparison to estimates from other countries, but since the analysis was restricted to the 30-<90 age range the results would not be comparable anyway, even using the WHO standard population. Using a NZ standard population has the benefit of giving more realistic incidence rates, as the NZ population is older than the WHO population.

For the initial exploration of whether age-specific rates of CRC were proportional between periods and between cohorts, which would indicate if the data follow age-period or age-cohort pattern, the following four classical plots for all CRCs were constructed (Carstensen, 2007):

1. Rates across age brackets where lines connect observations within each period
2. Rates across age brackets where lines connect observations within each birth cohort
3. Rates across period brackets where lines connect age groups
4. Rates across cohort brackets where lines connect age groups

These plots were based on incidence data tabulated in 5-year brackets, as the brackets need to be large to produce sufficiently stable rates. The rates were plotted on a log scale; therefore if the rates are proportional between periods (i.e. follow an age-period model) the lines in plot 1 and 3 will be parallel, and if the rates are proportional between cohorts, plot 2 and plot 4 will show parallel lines (Carstensen, 2007). In addition, a visual display of the three time scales, for interpretation of APC data along these three time scales, was presented using a hexamap as proposed by Jalal and Burke (2019).

3.2.2.2 Statistical model

The APC analysis aims to determine the contributions of age, period and cohort effects to the CRC incidence rates. The age effect represents incidence rates of CRC in different ages, the cohort effects show changes in the incidence rates across groups of individuals defined by birth cohort, while the period effects represent changes in the rates over time that influence all age groups simultaneously.

Using APC analysis, the incidence of CRC diagnosis was modelled using Poisson regression with Lexis triangles as observational units. To investigate patterns in CRC incidence rates (IRs) the following additive APC model for the log(IR) was fitted using the *apc.fit* function from the *Epi* package:

$$log[\lambda(a, c, p)] = f(a) + h(c) + g(p) \tag{3.9}$$

where: $f(a)$ is the log(IR) for age group a in the reference cohort adjusted for period deviations, with the global intercept included in $f(a)$,
$h(c)$ is the log(IRR) for cohort c relative to the reference cohort,
adjusted for age effect and period deviations with the net drift allocated to $h(c)$,
$g(p)$ is the log(IRR) for period p relative to the reference period,
adjusted for age and cohort effects.

Birth cohort 1946.5 corresponding to the middle cohort was chosen as the reference, and the middle period 2006.5 as the reference period.

The following estimable functions were determined using the maximum likelihood estimator (MLE):

1. The net drift, that is the overall log-linear trend by calendar year equivalent to the overall average annual percentage change (AAPC) in incidence rates across the 25-year period. The net drift was extracted from the fitted APC model, weighting the Lexis triangles according to the number of CRC cases.

2. Local drifts, that is the age-specific log-linear trends in IR by calendar year equivalent to AAPC.

3. The longitudinal age curve which indicates the fitted longitudinal age-specific incidence rates (IR) in the reference cohort adjusted for period deviations, presented per 100 000 person-years.

4. The cohort effect as IRRs relative to the reference cohort, adjusted for age and period deviations.

5. The period deviations as IRRs relative to the reference period adjusted for age and cohort effects.

Age, period and cohort effects were modelled using natural splines, following recommendations by Carstensen (2007) and used e.g. in Rutherford et al. (2010); Beal et al. (2018); Araghi et al. (2019); Dobson et al. (2020). Because the choice of the number of knots and their location has influence on the estimated age, period and cohort effects, I aimed at choosing the optimal number and location of the knots. Following the principle of choosing a model which is, on the one hand, complex enough to capture meaningful trends in the data (avoiding under-fitting) and, on the other hand, simple enough to avoid fitting

meaningless random fluctuations (avoiding overfitting) (Simpson, 2018b), the choice of the number and location of the knots for the splines was made by comparison of the goodness-of-fit (explained in the next Section 3.2.2.3) for two models fitted using the following strategies:

1. An initial model including knots located at every second year on each of the three time-scales was fitted. Using backwards elimination based on the Bayesian Information Criterion (BIC), the model was simplified under the constraint of having a minimum of four knots on each of the three time-scales to make sure that any statistically significant deviations were retained.

2. In the second model, the number of knots for each time-scale was the same as chosen by the backwards elimination (to allow a fair comparison), however with the knots' location based on the default approach used in the *Epi* package as explained by Carstensen (2007); knots on the age, period and cohort time-scales for the splines were placed in such a way that the number of CRC cases between each pair of successive knots is N/n, and N/2n below the first and above the last knot (where n is the number of knots and N is the total number of CRC cases).

To determine whether accounting for period differences achieved a statistically significant improvement in model fit, the fitted age-period-cohort models were compared to age-cohort models using a likelihood ratio (LR) test as implemented in the *apc.fit* function (Carstensen, 2007). Similarly, investigation of the improvement in the model fit of the age-period-cohort over the age-period model was carried out. The hypotheses that all cohort and period deviations were equal to zero were tested. The net drift, extracted using weighting by the number of CRC cases in each Lexis cell, was provided by the output from the APC model. The local drifts were not provided by the *apc.fit* function and were, therefore, estimated using a different implementation of the APC model provided in Carstensen (2019).

A significance level of 0.05 was used throughout and 95% CIs were reported where relevant. Data were analysed using R version 3.6.2. for OS X.

3.2.2.3 Model choice and validation

To validate the two models described above (points 2 and 1 in section 3.2.2.2), and to decide which of the two strategies for knot selection gives a model which fits the data better, a goodness-of-fit analysis was carried out following Sung et al. (2019). The observed age-specific incidence rates for each birth cohort (in 5-year groups) were compared to the 95% CIs corresponding to the fitted rates in those groups. Further, the overall agreement between the observed and predicted number of CRC cases in each age-cohort group was tested using Pearson's Chi-Square test. The model which provided better fit to the data (assessed by comparison of p-values) was used for all statistical analyses.

3.2.2.4 Sub-group analyses

Sub-group analyses were carried out to investigate trends in cancer incidence and to assess IRs in the following population strata: males, females, Māori and non-Māori. For males and females, sub-site specific analyses were carried out, but not for Māori and non-Māori, due to insufficient statistical power. For each strata, the same model chosen for the main APC analysis (that is, using the same knots and the same parametrisation) was fitted using Lexis cells only for this specific stratum. Fitting separate models for population strata was necessary, as the *Epi* package does not support covariates other than the time-scales. The comparisons of net drifts between sub-groups was based on a Z-test.

In the same way as the presentation of the results for the main model, cohort- and period-adjusted, sub-site- and sex-specific IR curves (for the age effect) and IRR curves

(for the cohort and period effects) were plotted, and presented in one graph. In the sub-site specific analyses, data from patients with multiple tumours diagnosed in different anatomical sub-sites at the same time were used in analysis for each anatomical location. The net drifts and the local drifts by 1-year age groups, all with 95% CI, were also estimated by fitting the APC model to subsets of the population.

To compare the age effects between Māori and non-Māori, cumulative incidence rates from age 30 to 90 years for each ethnicity were calculated based on longitudinal age effects (adjusted for cohort and period effects) . The curves for Māori and non-Māori were plotted and compared visually. The predicted median ages based on cumulative incidence rates were calculated using the *approx* function in R. The cumulative incidence rates were normalised by dividing the cumulative incidence by the cumulative incidence at age 89.5 following. Such normalised curves can answer the question whether Māori get CRC at a younger age than non-Māori, independently of the cohort effect and population structure, and the difference can be quantified.

Due to the undercounting of Māori CRC patients in NZCR before year 2006, all analyses stratified by ethnicity were repeated using Lexis cells with corrected Māori and non-Māori counts as explained in section 3.2.1.2. Results of the analysis with uncorrected counts are presented as the main results, including: descriptive statistics; ASRs to the structure of the NZ 2018 population for 1-year period brackets for 30-<90 years old, and for the population younger than 50 years; results of the APC models fitted separately for Māori and non-Māori (models checked for goodness of fit); model predicted IRs for both ethnicities; and plots with cumulative longitudinal and cross-sectional age effects.

The decision to base the main analysis on uncorrected counts was supported by my chief supervisor and was based on the following reasoning: the decrease in the size of the correction factor from year 2004 to year 2006 is dramatic, from 1.31 to 1.01 (for years 1997–2000 the correction factor is as high as 1.48) and such a big change does not seem

to be plausible in such a short time. The records of ethnicity in NZCR since 1999 are based on NHI recordings that can be several years old (Shaw et al., 2009), thus, rather a smooth change would be expected as the old NHI records were gradually replaced by new records. The use of the correction factors seems to give unrealistic ASRs thus the results of APC analyses can be highly influenced by the high correction factors. As explained by Boyd et al. (2017), in general, there is no methods for estimating the quality of the linkage used for producing the correction factors and therefore it is difficult to decide which results are more reliable. Considering those issues, both analysis were carried out and presented to show the differences, and allow the reader to judge the problems with reporting on CRC incidence in Māori before 2006. The results using corrected counts are presented as sensitivity analysis explained in Section 3.2.4.1.

3.2.3 Presentation of results of the APC analysis for use in clinical practice

In order to show how the predictions given by the APC model can be communicated to health care professionals and to policy-makers, predicted incidence rates for period 2018 were plotted as a function of age. These predictions were presented for proximal, distal and rectal cancers, stratified by gender. Such curves could be constructed for any population strata (e.g. for Māori females), subject to data availability. Period 2018 was used for those curves as in clinical practice the most recent data are the most relevant for decision making. For each curve, the data points were calculated by multiplying the age, period and cohort coefficients from the fitted model, for each Lexis cell corresponding to the date of diagnosis 2018.666 (i.e. 1 September 2018).

Additionally, although historical data are not relevant for decision making, age curves for the periods 1994, 2002, 2010 and 2018 were constructed for Māori and non-Māori to illustrate how the age-specific IRs changed over time for those ethnic groups. The

predictions from the APC models were calculated using R code provided by Carstensen (2011), however, I did not come across any study which used APC models to analyse CRC incidence data and used this presentation.

3.2.4 Sensitivity analyses

3.2.4.1 Sensitivity analysis for the undercount of Māori in NZCR

There are different methods that can be used for correction of the underestimated number of Māori in NZCR before year 2006, such as probabilistic linkage as implemented e.g., by Shah et al. (2012) which however would not be achievable for this PhD project. I used a method proposed by Shaw et al. (2009) for use in studies analysing cancer incidence data with census data as denominator. The method uses correction factors provided by the authors (specified in section 3.2.1.2).

To investigate the sensitivity of the analysis by ethnicity to the correction of the undercount of Māori CRC patients in NZCR before 2006, the following analyses were carried out using Lexis cells with corrected counts of CRC cases for both ethnicities: estimation of ASRs; fitting APC models separately for Māori and non-Māori (model check for goodness-of-fit); and calculation of cumulative longitudinal age effects. In the sensitivity analysis, the same APC models with respect to the location of knots and parametrisation as in the analysis with uncorrected counts were fitted. The same presentation of the results is provided as for the analyses with uncorrected counts.

3.2.4.2 Sensitivity analysis for the drift allocation

To investigate how much the choice of allocation of the drift to cohort effect influenced the interpretation of the results, I carried out an analysis using the same model as in the

main analysis, but with the cohort slope set to zero and the drift allocated to period (the reference cohort and period were the same as in the main analysis).

3.3 Results

3.3.1 Description of the study population

Characteristics of the study population are provided in Table 3.1. Using data from the NZCR for the period 1994 to 2018, a total of 65,530 incident CRC patients 30-<90 years old were registered among 59,723,000 person-years. Of the CRC patients, 52.1% were males, 5.3% were Māori, 66.5% had colon tumours, 31.7% had tumours in the rectum, and 1.8% had tumours in both locations. The median age at diagnosis in males was 70 years (IQR; 62, 77), in females 72 years (IQR; 62, 79), in Māori 64 years (IQR; 55, 72), and in non-Māori 71 years (IQR; 62, 78). In all subsequent 5-year periods there was an increase in the number of patients diagnosed with CRC. For proximal, distal and rectal tumours the number of registrations also increased in the subsequent periods. The increasing numbers of registrations with increasing period reflects the growth and ageing of the population, rather than increasing incidence rates, as the age-standardised rates actually decreased during the study period (Figure 3.2). The median age of the NZ population (30-<90 years) increased from 47 years (IQR; 38, 62) in 1994 to 52 years (IQR; 41, 65) in 2018. A higher proportion of patients younger than 60 years were diagnosed with tumours in the distal colon or rectum compared with patients 75-<90 years old, among whom proximal tumours predominate. Registrations of incident CRC by anatomical sub-site, cross tabulated against year of diagnosis (grouped in 5-year periods), age at diagnosis, gender, ethnicity and Duke's stage of cancer at diagnosis are summarised in Table 3.1.

Subset	Age at diagnosis		PYx1000	Total CRCs (%)	Sub-site		
	Median	IQR			Proximal colon (%)	Distal colon (%)	Rectum (%)
All	71	62, 78	59723	65530	23512 (37.5)	17817 (28.4)	21400 (34.1)
Period							
1994–1998	69	61, 77	10120	11588 (17.7)	3805 (16.2)	3129 (17.6)	3767 (17.6)
1999–2003	70	62, 78	11053	12434 (19.0)	4435 (18.9)	3386 (19.0)	4069 (19.0)
2004–2008	71	63, 78	12043	13000 (19.8)	4808 (20.4)	3458 (19.4)	4190 (19.6)
2009–2013	71	63, 79	12770	13902 (21.2)	5145 (21.9)	3816 (21.4)	4485 (21.0)
2014–2018	71	62, 79	13736	14606 (22.3)	5319 (22.6)	4028 (22.6)	4889 (22.8)
Age group							
<60	53	47, 57	41872	13026 (19.9)	3476 (14.8)	3790 (21.3)	5336 (24.9)
60-<75	68	64, 71	12472	28136 (42.9)	9338 (39.7)	8035 (45.1)	9755 (45.6)
75-<90	80	77, 84	5379	24368 (37.2)	10698 (45.5)	5992 (33.6)	6309 (29.5)
Gender							
Male	70	62, 77	28657	34158 (52.1)	9831 (41.8)	9930 (55.7)	13060 (61.0)
Female	72	62, 79	31065	31372 (47.9)	13681 (58.2)	7887 (44.3)	8340 (39.0)
Ethnicity*							
Māori	64	55, 72	6081	3452 (5.3)	928 (3.9)	964 (5.4)	1399 (6.5)
non-Māori	71	62, 78	53642	62078 (94.7)	22584 (96.1)	16853 (94.6)	20001 (93.5)
Duke's stage**							
I	71	63, 78	–	12836 (23.8)	4698 (23.8)	3912 (26.6)	4085 (23.2)
II	73	65, 79	–	8562 (15.9)	4269 (21.7)	2464 (16.8)	1736 (9.8)
III	70	62, 78	–	13590 (25.2)	5706 (29.0)	4010 (27.3)	3739 (21.2)
IV	69	60, 77	–	10921 (20.2)	3823 (19.4)	3095 (21.1)	2959 (16.8)
Unknown	72	62, 81	–	8380 (15.5)	1348 (6.8)	1304 (8.9)	5211 (29.6)

*Based on uncorrected counts of Māori and non-Māori CRC cases

**Not including cases for the period 1994–1998

Table 3.1: *CRC registrations in NZ for years 1994-2018, 30-<90 years old, for 5-year periods of diagnosis, age groups, gender, ethnicity and Duke's stage, by anatomical sub-site. The table also included the number of person-years in the NZ population 30-<90 years.*

3.3.2 Description of CRC incidence rates

3.3.2.1 Age-standardised rates

The overall age-standardised rates of CRC incidence decreased from 1994 to 2018. This decrease was similar for cancers located in the proximal colon, the distal colon and in the rectum. The incidence rates decreased similarly in both genders, however the trends differed between ethnicities; in non-Māori the trend was similar to that observed in the general population (i.e. to the overall trends), but in Māori the rates remained fairly stable during the 25-year study period (Figure 3.2).

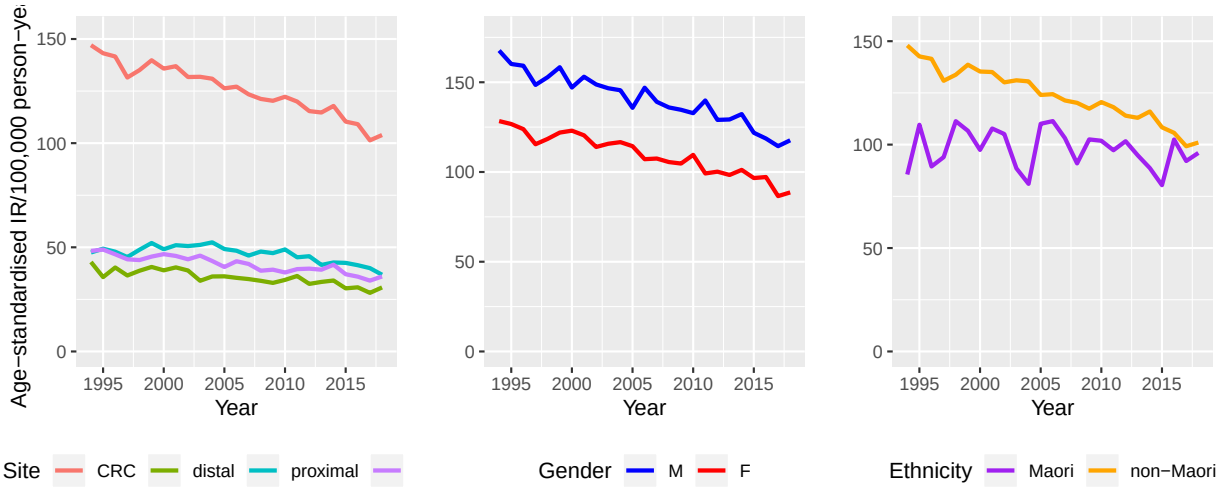


Figure 3.2: *Trends in CRC incidence rates age-standardised to the 2018 NZ population for age 30-<90 years for 1-year periods, by anatomical location of tumours, gender and ethnicity. The ASRs by ethnicity are based on uncorrected counts of Māori and non-Māori CRC cases.*

The age-standardised rates, however, have been decreasing only for individuals 50-<90 years old. The age-standardised incidence rates for young adults 30-<50 years old have been increasing since approximately the year 2007 (Figure 3.3).

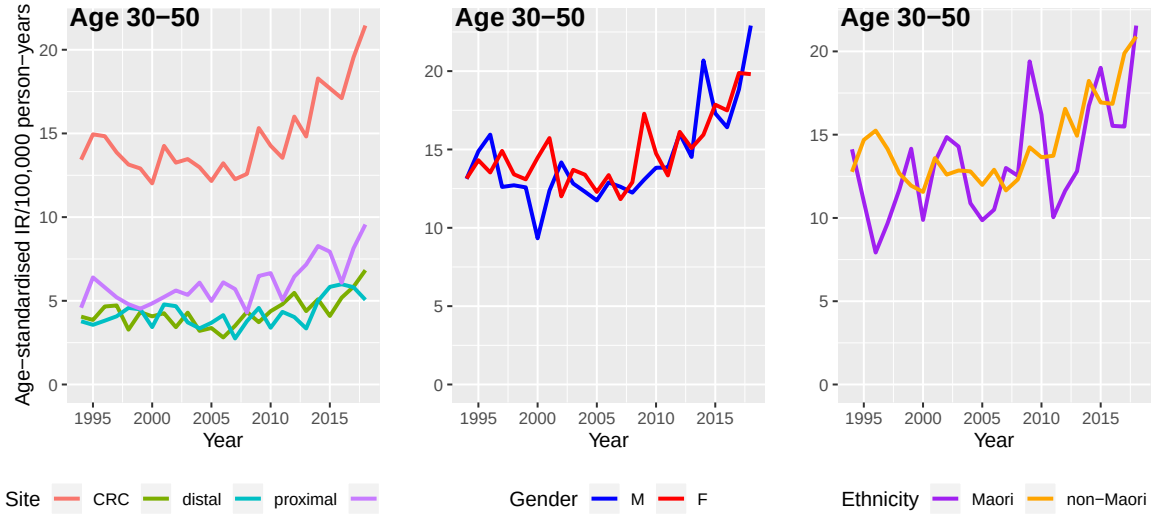


Figure 3.3: *Trends in CRC incidence rates age-standardised to the 2018 NZ population for age 30-<50 years for 1-year periods, by anatomical location of tumours, gender and ethnicity. The ASRs by ethnicity are based on uncorrected counts of Māori and non-Māori CRC cases.*

3.3.2.2 Age-specific rates

Age-specific rates based on the numbers of CRC cases and person-years for 5-year periods and 5-year age brackets, given in Table 3.2, were used to plot the four classical graphs presented in Figure 3.4. Plot A (in Figure 3.4) shows the crude rates as a function of age for each 5-year period, and reveals that for the younger adults, 30-<40 years, the IR was higher in the last 5-year period (2014–2018) compared to earlier periods, while for individuals aged 60-<75 years, rates were much higher in the earlier periods. Plot C shows the same pattern, as well as a fairly flat trend for those aged 45-<55 and stable rates for the oldest groups. The curves in plots A and C are not parallel, which indicates interactions between age and period present in the data and implies that the data do not follow an age-period model. Inspection of plots B and D suggests that cohort effect might explain the age-period interactions visible in plots A and C. However, this can only be assessed by investigating whether an APC model provides good fit to the data, as the assessment of whether the curves are parallel in plot B and in plot D is difficult

due to the small overlap between the age groups in non-successive 5-year cohorts (plot D) and a small overlap between cohorts in non-successive age groups (plot B).

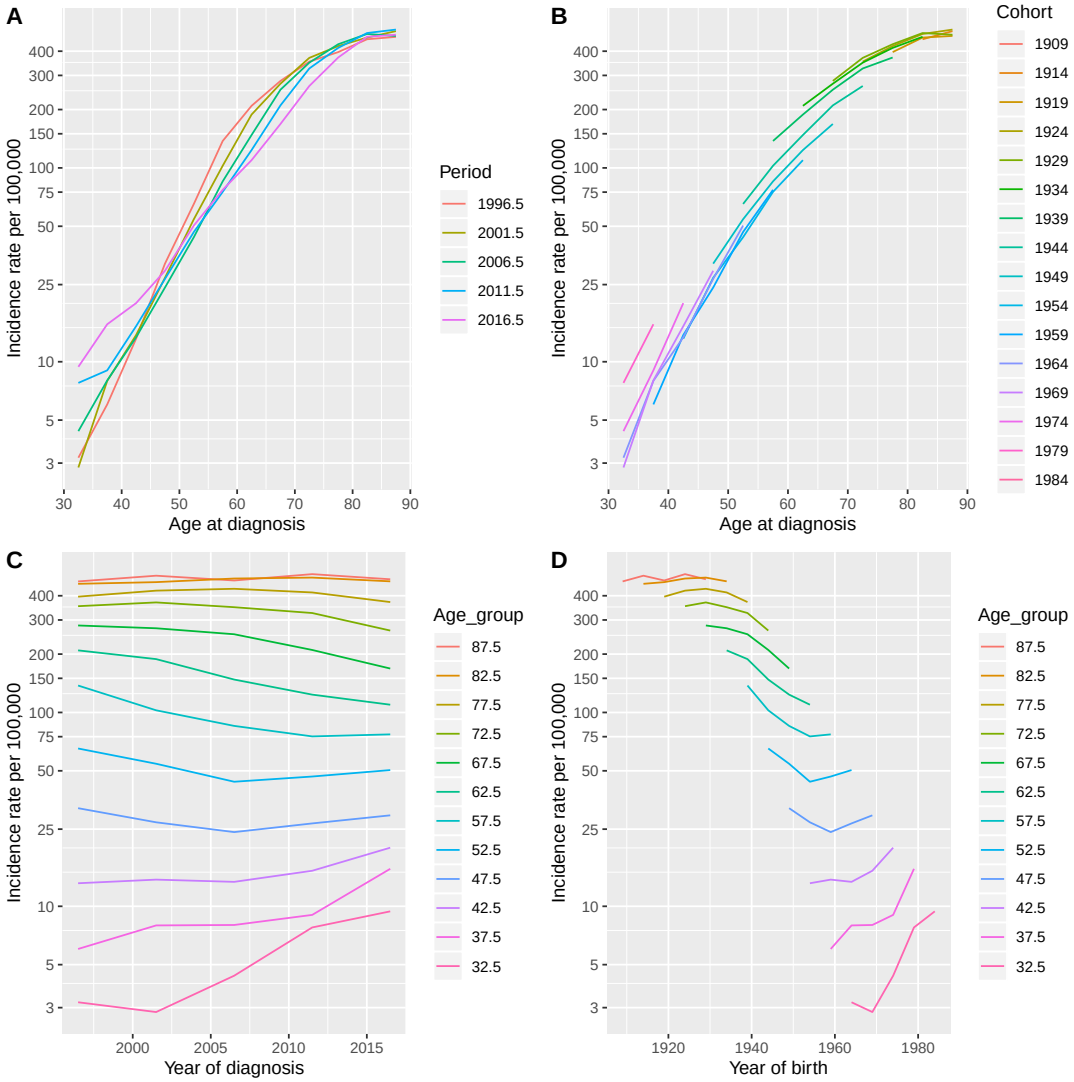


Figure 3.4: *Classical plots for the CRC incidence rates based on the Lexis diagram. The rates are for CRC registrations in NZ, for 30-<90 years old, for years 1994–2018 tabulated by 5-year age groups and 5-year periods (indicated by the range midpoints). The cohorts were calculated as period range midpoint minus age range midpoint (e.g. 1909 corresponds to birth cohorts 1904–1913. The 5 year overlap between successive cohorts is due to the use of Lexis squares for the classical plots. Plot A: age-specific rates by period of diagnosis. Plot B: age-specific rates by date of birth. Plot C: period-specific rates by age. Plot D: cohort-specific rates by age.*

	CRC cases					Person-years x 1000					Incidence rates/100000 person years				
	Mean period					Mean period					Mean period				
Mean age	1996.5	2001.5	2006.5	2011.5	2016.5	1996.5	2001.5	2006.5	2011.5	2016.5	1996.5	2001.5	2006.5	2011.5	2016.5
32.5	48	42	62	105	144	1501	1475	1414	1350	1529	3.20	2.85	4.38	7.78	9.42
37.5	90	123	125	131	225	1494	1544	1561	1451	1442	6.02	7.97	8.01	9.03	15.61
42.5	176	208	214	241	303	1339	1515	1601	1580	1510	13.15	13.73	13.37	15.25	20.06
47.5	397	363	372	426	472	1238	1339	1541	1592	1602	32.07	27.12	24.14	26.76	29.46
52.5	660	664	587	707	797	1013	1221	1337	1514	1580	65.17	54.37	43.91	46.71	50.43
57.5	1152	1021	1028	988	1155	837	995	1206	1313	1499	137.69	102.57	85.23	75.23	77.07
62.5	1466	1541	1446	1465	1432	701	818	978	1185	1306	209.07	188.45	147.81	123.58	109.64
67.5	1885	1813	1994	1989	1955	671	667	788	947	1160	281.00	271.79	253.15	210.06	168.48
72.5	2054	2223	2138	2378	2357	581	602	612	730	891	353.39	369.51	349.18	325.76	264.52
77.5	1687	2067	2243	2217	2410	426	486	516	533	650	395.72	425.03	434.43	415.98	371.05
82.5	1301	1496	1833	2010	2038	282	318	373	405	429	460.87	470.23	490.87	496.22	474.63
87.5	672	873	958	1245	1318	142	172	200	241	271	474.17	507.53	479.53	517.11	486.45

Table 3.2: *Number of CRC cases and person-years in NZ population 1994–2018 in 5-year age groups and 5-year periods, indicated by the range midpoints.*

The hexamap in Figure 3.5 gives a different visualisation of the patterns seen in the classical plots in Figure 3.4. The isolines for 30-<45 years old show increasing incidence rates with increasing period, changing from dark blue to light blue at the bottom of the diagram. For ages 45-<55 years old the incidence rates are fairly stable, for intermediate ages (55-<75 years) the diagram shows decreasing incidence rates with later periods and stable rates for the oldest individuals.

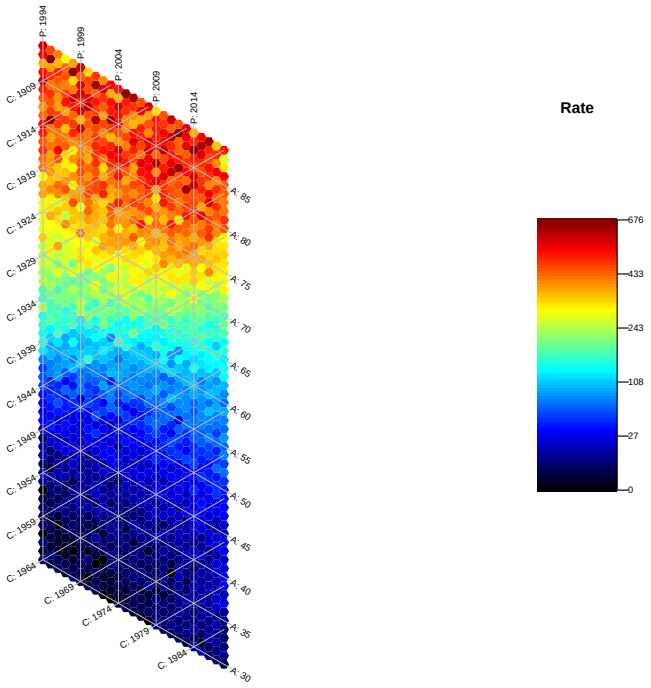


Figure 3.5: *Hexamap showing patterns in the CRC incidence data in NZ from 1994 to 2018 as a function of age, cohort and period.*

3.3.3 Results of age-period-cohort modelling

This section provides results for several fitted APC models, and is structured as follows: firstly, I provide results of the goodness-of-fit analysis which informed the choice of the model used in the main APC analysis and in all subgroup analyses; secondly, I present

the results from the fitted APC model for the overall CRC incidence and the results from subgroup analyses; thirdly, I give the result of sensitivity analysis; and finally I provide model-based incidence rates as an example of the presentation of the results for use in clinical practice.

3.3.3.1 Model choice and model validity

Goodness-of-fit analysis showed that of the two models tested (one based on the default knot selection from the *Epi* package and one using backwards elimination), only the model with knots selected by the backwards elimination procedure did not show disagreement between the predicted and observed IRs (p=0.20), which is shown in Figure 3.6. The figure shows 95% CIs for the IRs predicted by the APC model for 5-year age groups and 5-year cohort groups, compared to the observed values.

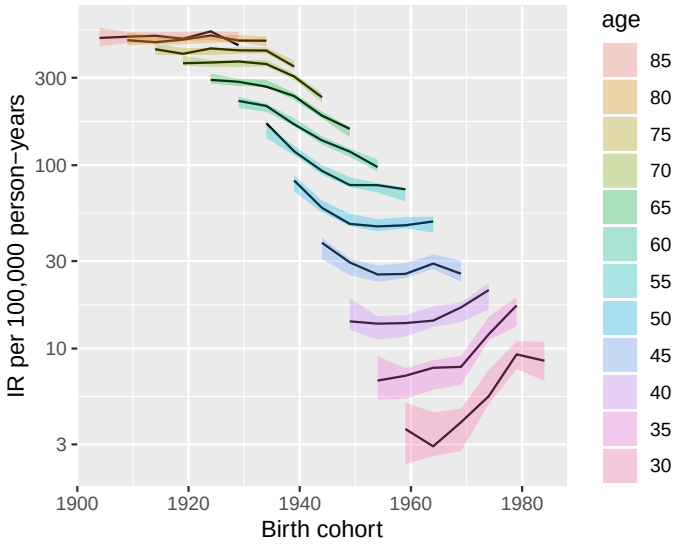


Figure 3.6: *Goodness-of-fit for the model with knots based on backwards elimination. Age-specific incidence by birth cohort from 1904 (1904–1908) to 1984 (1984–1988). Lines denote observed incidence for 5-year age groups (e.g., 30 indicates lower age in group 30-<35) and shaded areas indicate 95% CIs for the corresponding fitted rates from the APC model.*

In contrast, the model which used the default procedure for the location of knots from the *Epi* package did not fit the data well ($p<0.001$) (Figure 3.7). As can be seen in Figure 3.7 the model fit to the data is particularly poor for the youngest age groups.

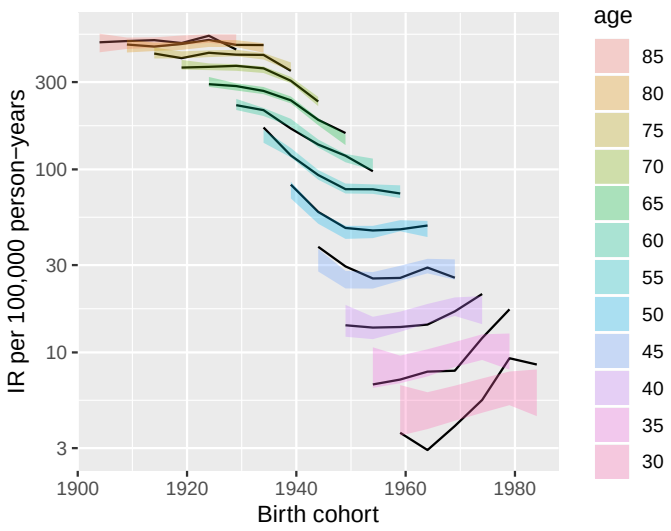


Figure 3.7: *Goodness-of-fit for the model based on the default approach from the Epi package. Age-specific incidence by birth cohort from 1904 (1904–1908) to 1984 (1984–88). Lines denote observed incidence for 5-year age groups (e.g., 30 indicates lower age in group 30-<35) and shaded areas indicate 95% CIs for the corresponding fitted rates from the APC model.*

Therefore, in all fitted APC models, the number and location of knots based on the backward elimination procedure were used. The knots were located as follows: seven knots located at ages 39, 57, 59, 77, 79, 81 and 83; eight knots for birth cohort located at years 1936, 1938, 1940, 1964, 1966, 1976, 1978 and 1980; and four knots for the period scale at years 2001, 2009, 2015 and 2017. All APC models were fitted using the following parametrisation: the global intercept was included in the age effect; period slope was assumed to be zero; drift was allocated to the birth cohort; birth cohort reference was 1946.5; and the period reference was 2006.5.

The goodness-of-fit analysis for each of the APC models fitted to subsets of the data did not identify any disagreement between the modelled and observed data. P-values for the

goodness-of-fit tests for all fitted models can be found in Table 3.6 in Section 3.3.3.4, which presents results for sub-group analysis.

3.3.3.2 Estimated age, period and cohort effects

The results of the APC analysis of the whole study population are shown in Figure 3.8. The contribution of age, period and cohort effects to CRC incidence rates are estimated under the assumption of zero period slope. The estimated age effect is presented as longitudinal age-specific IR for the reference cohort, adjusted for period and cohort effects (left pane); cohort effect is presented as IRR relative to the reference cohort adjusted for age and period (middle pane); and IRR for period deviations relative to the reference period adjusted for cohort and age are shown in the right pane.

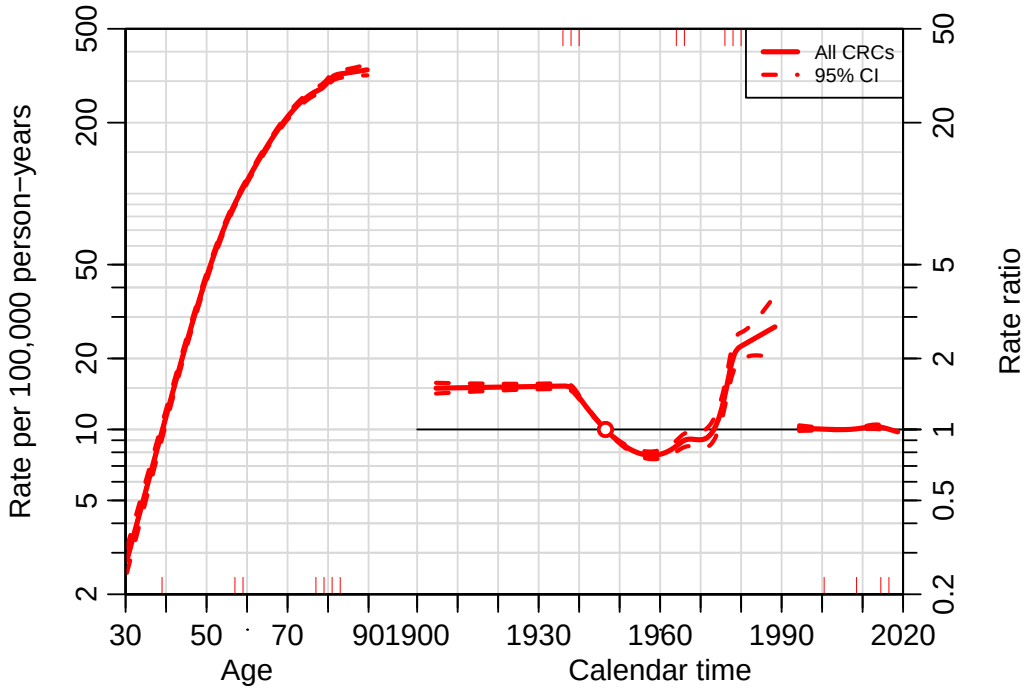


Figure 3.8: *The estimated age, period and cohort effects from the APC model for incident CRC, assuming zero period slope. The reference is the 1946.5 cohort to which the age-effect refers. Knots are indicated as short vertical red lines at the top and bottom of the diagram*

Compared to the reference cohort, the IRRs for cohorts born between 1904 and 1939 were increased by approximately 50%. In the subsequent cohorts born between 1939 and 1957, there was a sharp decrease in incidence rates. The IRR in those born in 1939 was over 80% higher than in individuals born in 1957 when the IRR was the lowest. The IRR started to increase slowly in generations born in the late 60's, however the most dramatic increase affected those born between 1972 and 1988. The IRRs from the middle pane in Figure 3.8 for selected birth cohorts are shown in Table 3.3.

Cohort	IRR (95% CI)	Cohort	IRR (95% CI)
1904	1.49 (1.42, 1.57)	1949	0.89 (0.88, 0.90)
1909	1.50 (1.43, 1.57)	1954	0.79 (0.77, 0.82)
1914	1.50 (1.45, 1.56)	1959	0.78 (0.75, 0.81)
1919	1.51 (1.46, 1.56)	1964	0.87 (0.82, 0.92)
1924	1.52 (1.47, 1.56)	1969	0.90 (0.84, 0.97)
1929	1.52 (1.48, 1.56)	1974	1.13 (1.02, 1.26)
1934	1.53 (1.49, 1.57)	1979	2.22 (1.94, 2.54)
1939	1.39 (1.37, 1.42)	1984	2.46 (2.02, 2.99)
1944	1.08 (1.08, 1.09)	1988	2.65 (1.95, 3.59)

Table 3.3: *Incidence rate ratios for selected birth cohorts for incidence rates for CRC, relative to 1946.5 reference cohort, assuming zero period slope.*

The statistical significance of the components of the fitted APC model (Figure 3.8) are in the Anova table (Table 3.4). The estimated net drift and cohort deviations are highly statistically significant as indicated by the p-values in the Anova table. The model also indicates period deviations statistically significantly different from zero. However, these period deviations (Table 3.4) were minimal, most likely clinically not important, and therefore (following Carstensen (2007)) do not influence my interpretation of the results. The very pronounced cohort deviations explained 67% of the deviance that can be explained by the cohort effect (net drift plus cohort deviations) and period deviations,

while the period deviations explained only 0.5%.

Model	Mod. df.	Mod. dev.	Test df.	Test dev.	Pr(>Chi)	Test dev/df	H0
Age	2993	4919.236	NA	NA	NA	NA	
Age-drift	2992	4381.885	1	537.3519	0.00000	537.3519	zero drift
Age-Cohort	2986	3242.786	6	1139.0986	0.00000	189.8498	Coh eff | drift
Age-Period-Cohort	2984	3234.571	2	8.2151	0.01644	4.1075	Per eff | Coh
Age-Period	2990	4366.000	6	1131.4295	0.00000	188.5716	Coh eff | Per
Age-drift	2992	4381.885	2	15.8842	0.00036	7.9421	Per eff | drift

Table 3.4: *Anova table showing the deviance explained by components of the APC model (drift, cohort deviations and period deviations) for the overall CRC incidence.*

Based on the net drift extracted from the APC model, the overall CRC incidence rates in NZ from years 1994 to 2018 decreased with the net AAPC of -1.31 % (95%CI; -1.42, -1.20). However, these trends differed between ages with respect to magnitude and direction (Figure 3.9 and Table 3.5). The analysis showed decreasing rates of CRC in individuals around 50–80 years, with the strongest decrease in those around 60 years old, but increasing incidence rates in young adults aged from 30 to around 45 years. The strongest increase was observed in individuals 30 to around 40 years old, with AAPC over 4%. The incidence rates were stable in the oldest age groups (80-<90 years).

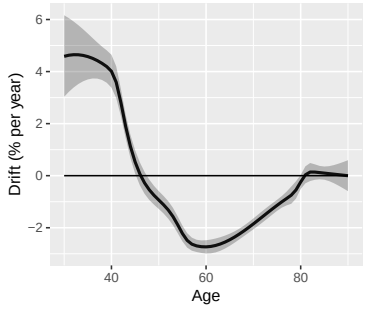


Figure 3.9: *Local drifts (equivalent to age-specific AAPC) for the overall CRC incidence from 1994 to 2018 in 1-year age groups.*

Age	Corresponding cohorts	AAPC	95% CI
30	1964–1988	4.59	(3.03, 6.16)
35	1959–1983	4.58	(3.70, 5.46)
40	1954–1978	4.01	(3.38, 4.64)
45	1949–1973	0.53	(0.08, 0.98)
50	1944–1968	-0.94	(-1.28, -0.59)
55	1939–1963	-2.22	(-2.52, -1.92)
60	1934–1959	-2.73	(-2.99, -2.48)
65	1929–1953	-2.46	(-2.70, -2.23)
70	1924–1949	-1.85	(-2.06, -1.64)
75	1919–1943	-1.14	(-1.35, -0.93)
80	1914–1939	-0.24	(-0.51, 0.03)
85	1909–1933	0.10	(-0.15, 0.35)
90	1904–1929	0.00	(-0.59, 0.60)

Table 3.5: *Local drifts showing the AAPC for years 1994–2018 for selected ages. The local drifts were estimated as a function of continuous age as shown in Figure 3.9. The table also indicates to which birth cohorts the specific drift applies.*

The differences in the magnitude and direction of the age-specific trends (local drifts in Figure 3.9) which are presented in Table 3.5 can be explained almost entirely by the very pronounced cohort deviations independent of the net drift (shown in Figure 3.10). Age-period interactions that could not be explained by the cohort effect might exist but were not detected by goodness-of-fit analysis (Section 3.3.3.1).

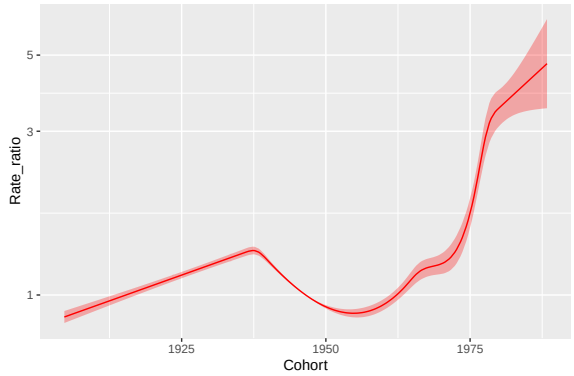


Figure 3.10: *Estimated cohort deviations for the overall CRC from the APC model with the cohort slope set to zero and the drift allocated to period (based on the model fitted in sensitivity analysis Section 3.2.4.2). The reference cohort is 1946.5.*

3.3.3.3 Model-based CRC incidence rates

Figure 3.11 shows the estimated incidence rates based on the fitted APC model for chosen ages, which are the model-based analogs of the crude age-specific rates shown in Figure 3.4 plot D. The top diagram shows the trends for selected ages between 55 and 90 years with a sharp decrease in incidence in generations born around 1939–1957. The decrease reflects the estimated cohort effect shown in Figure 3.8. The bottom graph shows trends for 30-<50 years old, with the increasing IRs in generations born around 1972–1988.

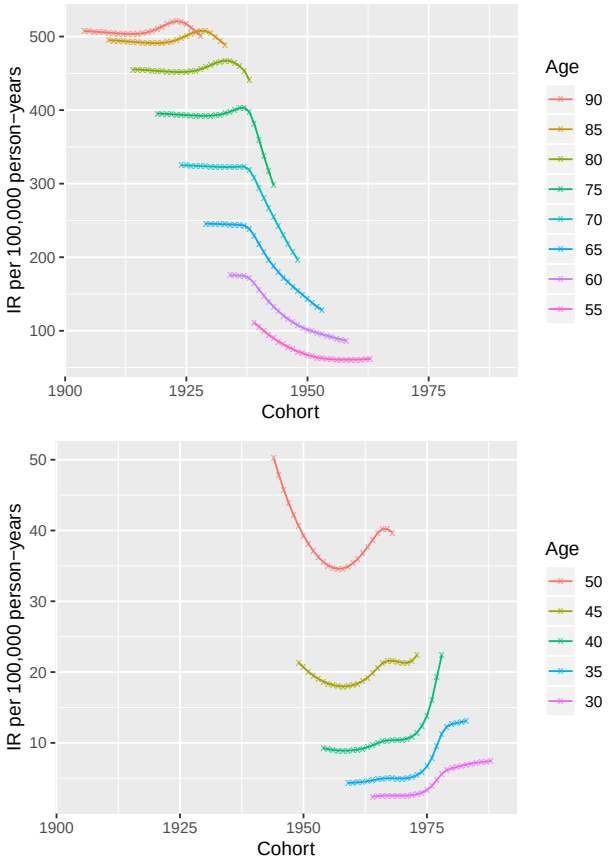


Figure 3.11: *Model-based incidence rates showing trends of CRC for selected ages. The y-axes are on linear scale in order to show the absolute increase in rates. Data are displayed in two graphs because of the difference in magnitude of IR between old and young groups.*

3.3.3.4 Sub-group analysis

This subsection begins the presentation of the results for sub-group analysis carried out for: males; females; Māori; non-Māori; and separately for males and females for each of the three anatomical sub-sites (proximal, distal and rectum). This is done by providing, for each of the fitted APC models: p-values for the goodness-of-fit test; the estimated net drifts extracted from the fitted APC models (equivalent to AAPC); the statistical significance of the difference in drifts between genders or ethnicities, and between males and females for each of the anatomical sub-sites (Table 3.6). I then show how those overall trends differ by age. Like the overall trends, the trends for sub-groups are always presented for the whole study period, from 1994 to 2018.

There was no disagreement in any of the models between fitted and observed values (assessed by chi-square test). The net drifts were negative for both genders, for non-Māori, and for all three anatomical sub-sites, showing the overall decrease in incidence over the studied period. However, in Māori there was no statistically significant change in CRC incidence rates, with the AAPC equal to 0.25% (95%CI; -0.24, 0.75). In sub-site specific analysis stratified by gender, the incidence of distal tumours decreased more rapidly in females than in males ($p<0.001$, Z-test). The opposite was observed in proximal tumours, with a stronger decrease in incidence in males than in females, although the difference was not statistically significant ($p=0.09$, Z-test). For rectal tumours a similar decrease in incidence rates was observed in both genders ($p=0.35$, Z-test).

In females, the decrease in incidence of distal tumours (located in the left part of the colon) was nearly 60% faster than in proximal tumours (located in the right colon), showing a growing proportion of right-sided colonic tumours over the period 1994–2018 ($p=0.004$, Z-test). In males, the opposite trend was observed. The decrease in proximal cancers was around 50% faster than in distal colonic tumours, showing an increasing proportion of distal tumours in males over time ($p=0.03$, Z-test). When

Subset	CRC count	AAPC (95% CI) (net drift)	P-value (model fit)
All	65,530	-1.31% (-1.42, -1.20)	0.20
Gender			
Male	34,158	-1.28% (-1.44, -1.13)	0.23
Female	31,372	-1.40% (-1.56, -1.24)	0.23
Ethnicity**			
Māori	3,452	0.25% (-0.24, 0.75)	0.35
non-Māori	62,078	-1.38% (-1.50, -1.27)	0.23
Proximal colon			
Male	9,831	-1.36% (-1.66, -1.06)	0.06
Female	13,681	-1.02% (-1.28, -0.76)	0.28
Distal colon*			
Male	9,930	-0.91% (-1.19, -0.62)	0.14
Female	7,880	-1.61% (-1.92, -1.30)	0.98
Rectum			
Male	13,060	-0.97% (-1.21, -0.73)	0.14
Female	8340	-1.16% (-1.46, -0.85)	0.18

*p<0.05, **p<0.001

Table 3.6: *Net drifts (AAPC) for fitted models and p-values for the goodness-of-fit. Stars indicate statistically significant differences in net drifts between genders or ethnicities (based on Z-test). The values for Māori and non-Māori are based on uncorrected counts of CRC cases.*

assessing the age-specific changes (i.e. local drifts expressed as AAPC) in CRC incidence rates for sub-groups by gender and by ethnicity, I found a similar change in males and females with respect to the magnitude and direction, but differences between ethnicities in some age groups (Figure 3.12). Based on a visual inspection of the curves, individuals 50-<75 years old, the rates in non-Māori decreased substantially [at age 60 years the AAPC was -2.96% (95% CI; -3.23, -2.70)], while the rates in Māori were fairly stable over the whole study period. However, in young adults, 30-<45 years old,

there was a strong increase in incidence rates of a similar magnitude in both ethnic groups.

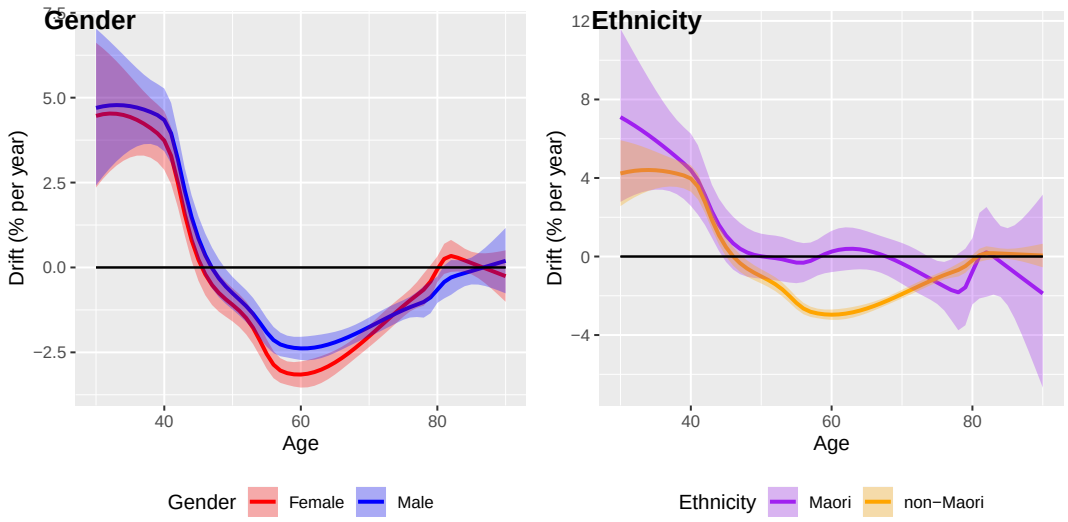


Figure 3.12: *Local drifts with 95% CIs by gender and by ethnicity (based on uncorrected counts of CRC cases).*

The differences in local drifts between ethnic groups can be explained by different cohort effects, shown in the middle panes in Figure 3.13. As can be seen in Figure 3.13, the shapes of the cohort effects in Māori were different than in non-Māori. The steep decrease in IRR in non-Māori born between circa 1939 and 1957 was not observed in Māori. However, the sharp increase in IRR in generations born from around the 1960s onwards, especially pronounced in generations born between the 1970s and 80s, affected both ethnic groups and was of a very similar magnitude.

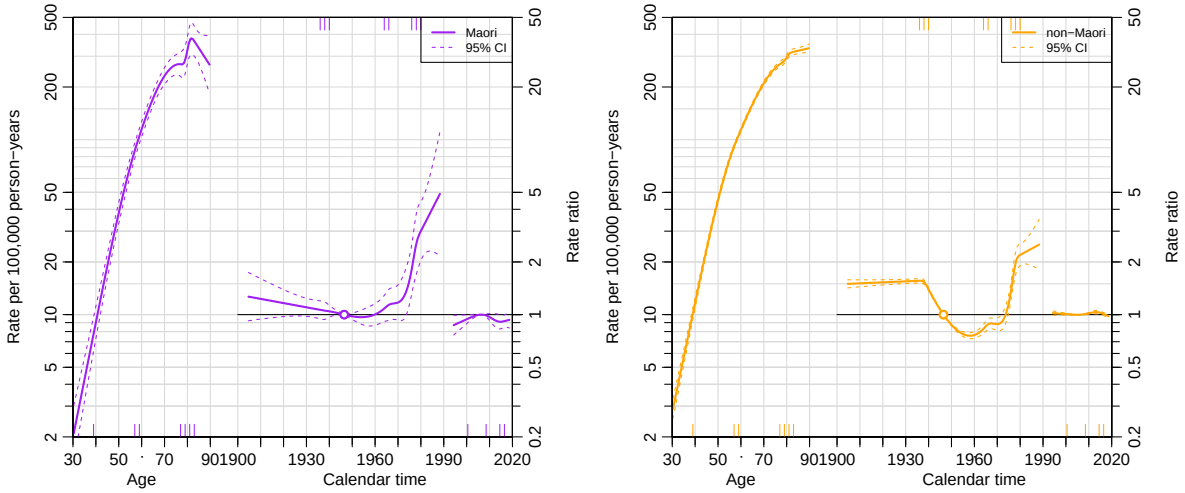


Figure 3.13: *The estimated age, period and cohort effects from the APC model for incident CRC by ethnicity, assuming zero period slope. The reference is the 1946.5 cohort to which the age-effect refers. The reference period is 2006.5. The results are based on uncorrected counts of CRC cases for Māori and non-Māori.*

The median age at diagnosis in Māori was 7.5 years lower than in non-Māori. The cross-sectional curves in Figure 3.14 show that a difference of around two years can be explained by a combination of age and cohort effects. The remaining 5.5 years can be explained by the population structure.

With respect to the two-year difference in median age that can be explained by age or cohort effects, the longitudinal cumulative incidence curves in Figure 3.15 show that only a small part (48 days) can be explained by age effect. The graph shows the distribution of the age at which, according to the model, individuals born in a specific year would be expected to be diagnosed with CRC, assuming that they will be diagnosed with CRC between 30 and 90 years of age. Based on the cumulative incidence, the median age at diagnosis would be 75.87 years for non-Māori and 75.74 years for Māori. Therefore, the remainder of the two-year difference can be explained by cohort effect.

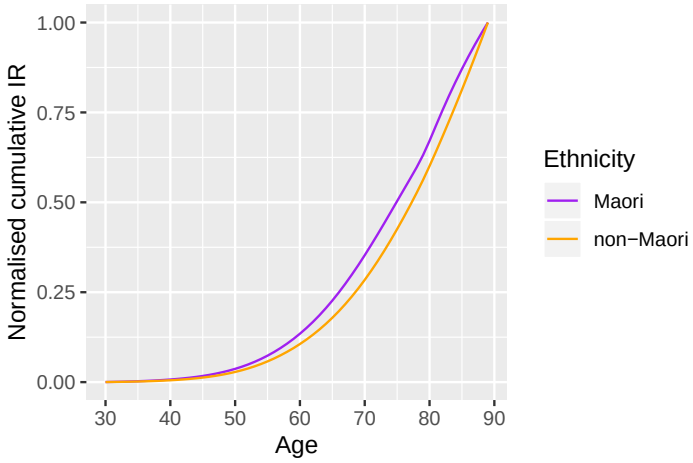


Figure 3.14: *Cross-sectional cumulative CRC incidence for registrations between 1994 and 2018 by ethnicity for ages 30-<90 years based on fitted APC models. The cumulative incidence rates (y axis) have been normalised to a cumulative incidence equal to 1 for age 90 years in order to compare the age effects in both ethnicities independently of the overall incidence.*

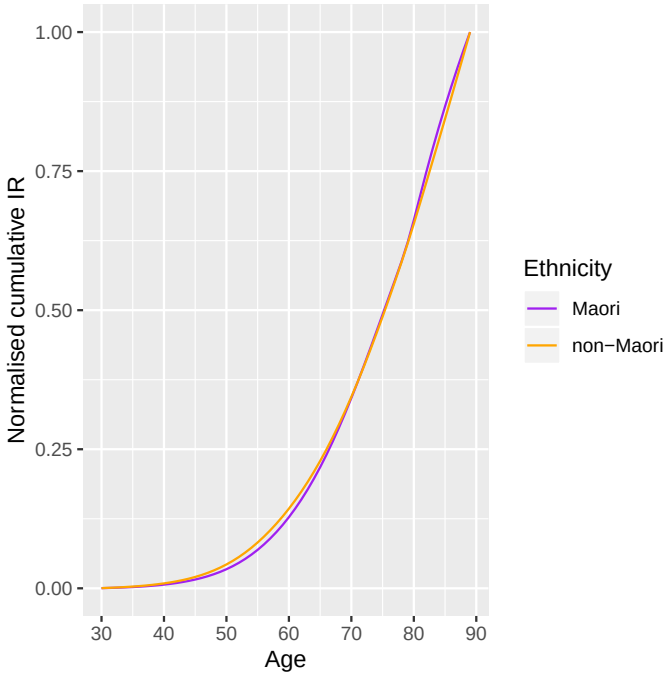


Figure 3.15: *Longitudinal cumulative CRC incidence for registrations between 1994 and 2018 by ethnicity for ages 30-<90 years based on fitted APC models. The cumulative incidence rates (y axis) have been normalised to a cumulative incidence equal to 1 for age 90 years in order to compare the age effects in both ethnicities independently of the overall incidence.*

With respect to the sub-site specific analyses, age-specific changes (local drifts) for proximal, distal and rectal tumours were different in males and in females, with the main differences in trends for proximal and rectal cancers (Figure 3.16). As can be seen, the incidence of proximal tumours increased in females around 78 years and older (the 95% CIs do not include the null value), while in males of the same age the increase was of a smaller magnitude and borderline statistically significant. Incidence of rectal tumours has been consistently decreasing in females 50 years and older, while in males older than 80 years the incidence was stable. In distal tumours an interesting drop in incidence, similar in males and females, was observed in the youngest patients studied, 30-<40 years old.

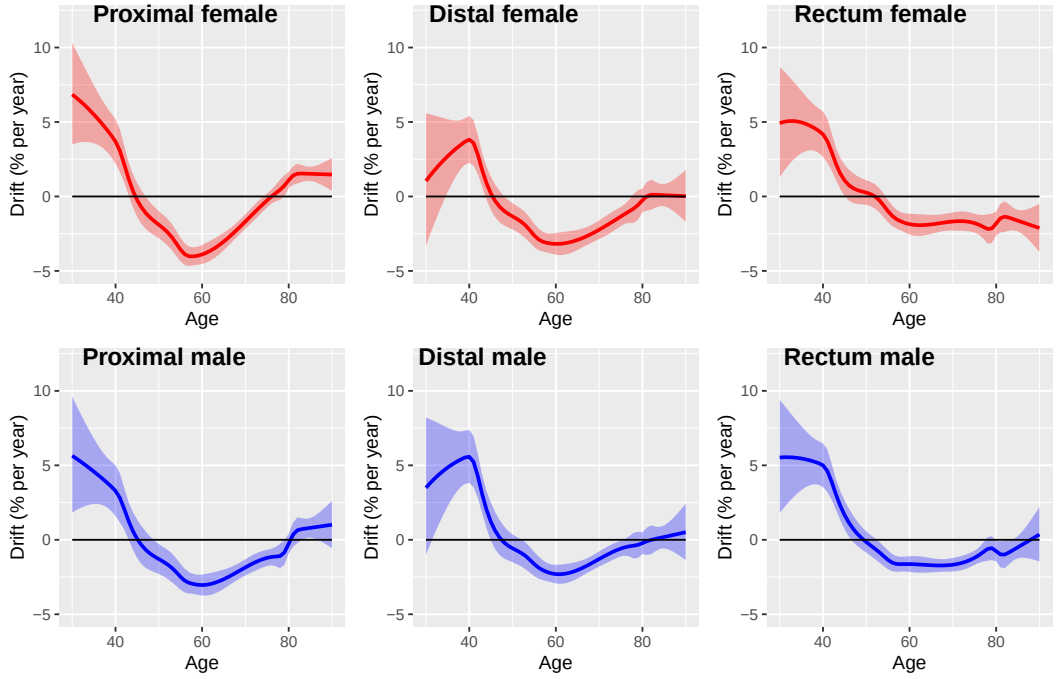


Figure 3.16: *Local drifts by gender and anatomical sub-site showing AAPC from 1994–2018.*

Across ages, in males, the direction and magnitude of the changes in rates were similar for all three sub-sites. In males younger then approximately 45 years the incidence rates increased substantially during the studied period but in older males, 50-<75 years old, the incidence rates decreased. In males older than 80 years the incidence rates in proximal cancers increased slightly during this period. In females younger than 75 years, the trends

were similar to the trends in males. In females younger than 45 years, the incidence rates in cancers located in all three sub-sites increased but in those 50-<75 years old the rates decreased. However, females 75 years and older experienced an increase in the incidence of proximal tumours of around 1.5% per year, but a decrease in rectal tumours of over 1.5% per year.

Figure 3.17 shows the estimated sub-site specific age, cohort and period effects for males and females that explain these age-specific trends. Although the patterns were similar for proximal and rectal cancers, with a steep increase in generations born from the 1970s onwards, the trend in distal tumours differed slightly from the trends in the other two sub-sites in both genders. The IRR in distal tumours, after the initial increase in generations born in the 1970s, decreased in those born in the 80s. The age effects differed between genders for all three sub-sites (the left panes). In females older than 75 years the incidence rates of proximal tumours is much higher than the incidence in the other two sub-sites. In males the proximal cancers were also associated with older age (in males older than 80 years) but the association was not as strong as in females.

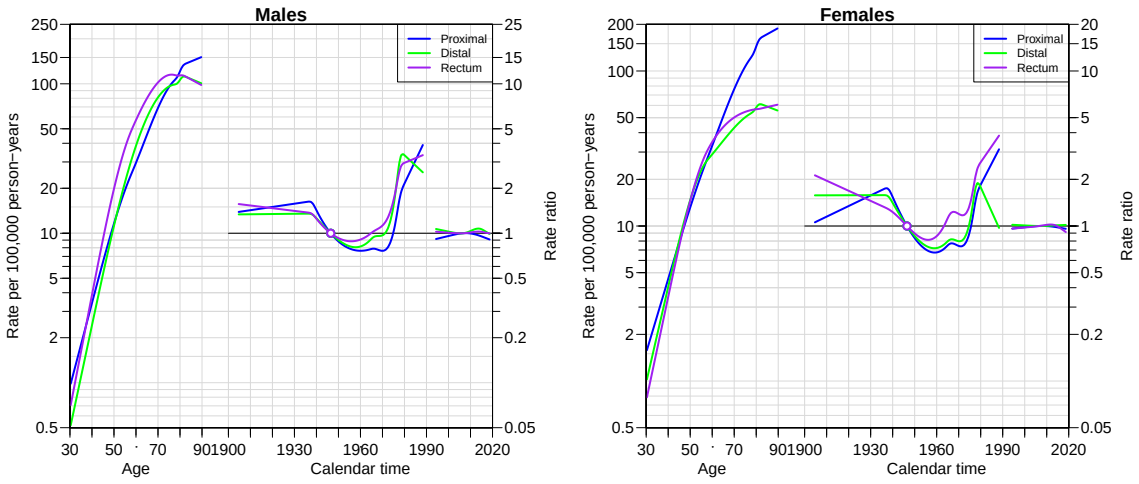


Figure 3.17: *The estimated age, period and cohort effects from the APC model for incident CRC by gender and anatomical sub-site, assuming zero period slope. The reference is the 1946.5 cohort to which the age effect refers. The reference period is 2006.5.*

Above I presented results of sub-group analysis for each of the three anatomical sub-sites, separately for males and for females. In clinical practice, those results might be important for the choice of investigation for presence/absence of CRC (i.e. colonoscopy, FS or digital examination). However, the estimated age, cohort and period effects for each sub-site by gender are quite complex. In order to make the results usable during a short medical consultation for assessment of Mary's CRC risk, in the next Section 3.3.4 I show examples of how the results from fitted APC models can be presented to doctors and policy-makers to help quickly assess the incidence rates in different population groups.

3.3.4 Presentation of the results for use in clinical practice

When considering the graphical presentation of the results of the APC analysis, the direct comparison of curves between strata is complicated by the fact that the curves change their relative location with a change of the parametrisation, e.g. a change of a reference category. In this section I therefore present curves based on predicted IRs for a single year, that allow direct comparison of the incidence rates between genders and between sub-sites. The comparison can be made within one graph or between different graphs constructed for different strata or for different periods of diagnosis. Figure 3.18 shows the model-based (predicted) incidence rates for proximal, distal and rectal cancer stratified by gender for calendar year 2018. For example, in 80-year-old females the IR for proximal tumours is nearly three times as high as the rates for rectal or distal tumours. The two bottom graphs in Figure 3.18 show the sub-site specific rates in young adults. In young males (30-<60 years), the rates of distal and rectal tumours are higher than in females of the same age but the incidence rates of proximal tumours are very similar.

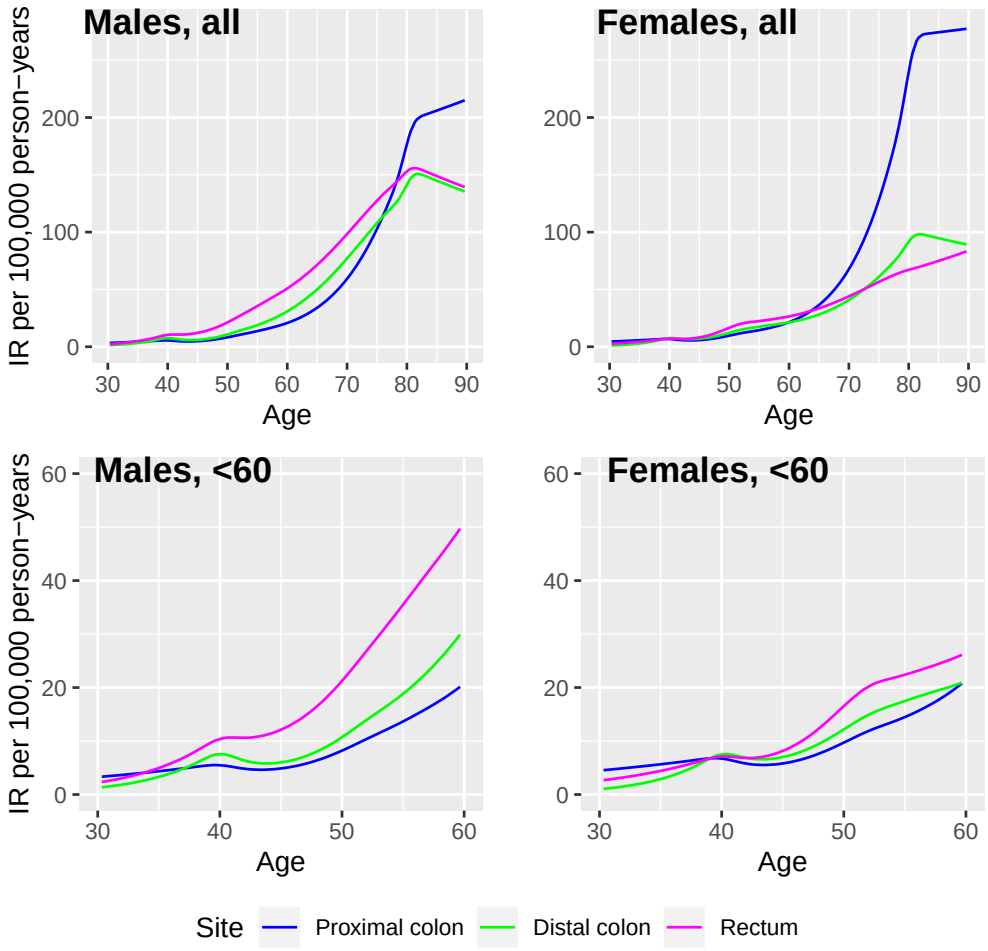


Figure 3.18: *Model-based, age-specific CRC incidence rates for anatomical sub-sites, stratified by gender for the year 2018. These charts are examples of how the results of the analysis can be communicated to health professionals. For this reason only point estimates are given, and the incidence rates are on a linear scale.*

Graphs for a direct comparison of incidence rates in Māori and non-Māori are shown in Figure 3.19. In addition to the year 2018 (the relevant year for use in clinical practice), there are three graphs for the years 1994, 2002 and 2010, to show how the incidence rates changed over the study period. In 1994 the rates in non-Māori were much higher than in Māori across the whole age range. In the following years, the incidence rates in non-Māori younger than 70 years gradually decreased to reach a lower level than in Māori of the same age. In the year 2018, in patients between 50 and 75 years old, the rates in

non-Māori had become lower than the rates in Māori. The decrease in rates in non-Māori is a reflection of the cohort effect which affected non-Māori and was shown in Figure 3.13.

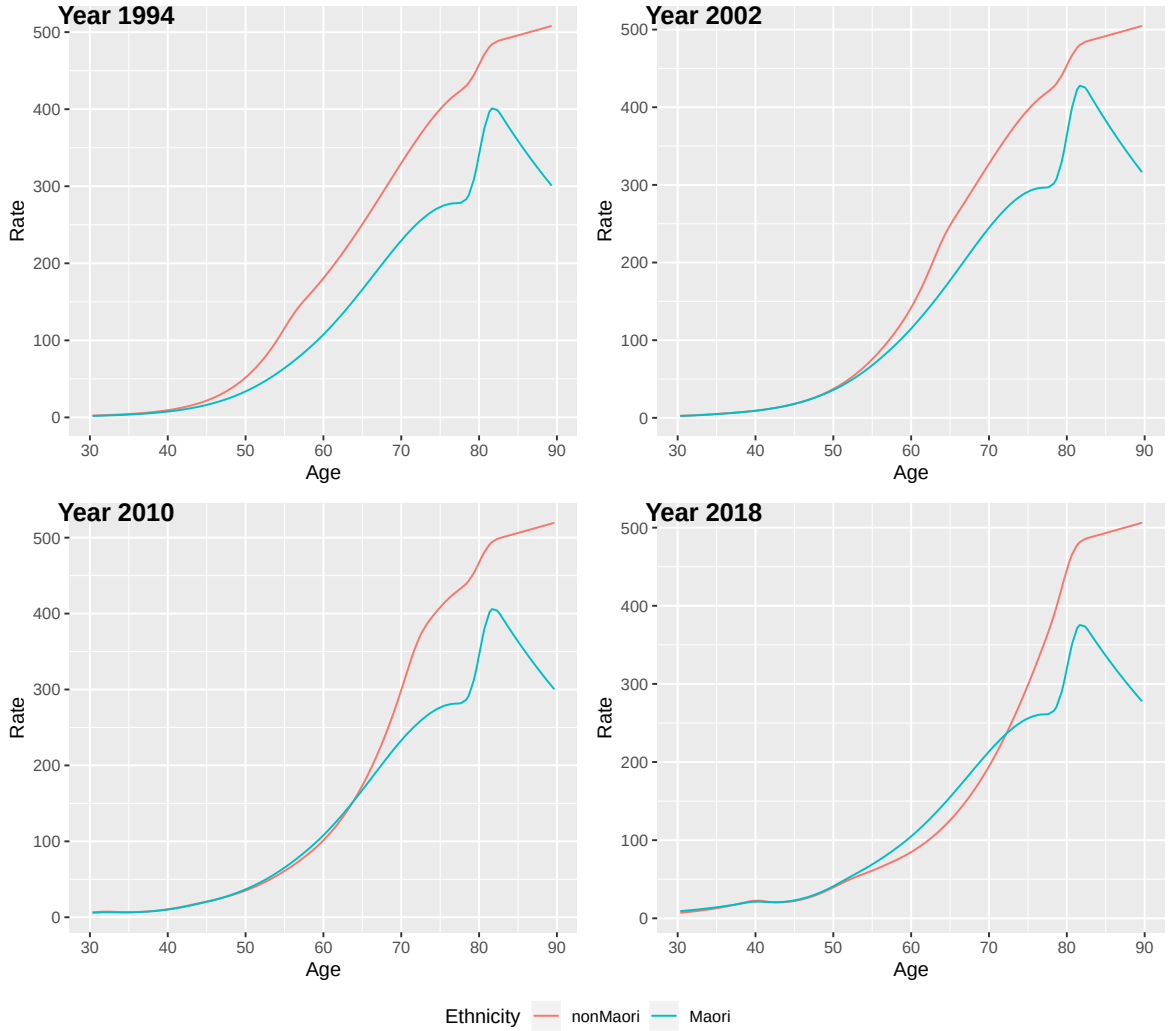


Figure 3.19: *Model-based age-specific CRC incidence rates by ethnicity, estimated by the APC model for the four chosen periods (based on uncorrected counts of CRC cases).*

As can be seen in Figure 3.20 which shows predicted IRs by ethnicity for 2018 with 95% CI, in the age range approximately 50–65 years, the IRs in Māori are higher than in non-Māori (the 95% CIs for the two ethnicities in that age range do not overlap).

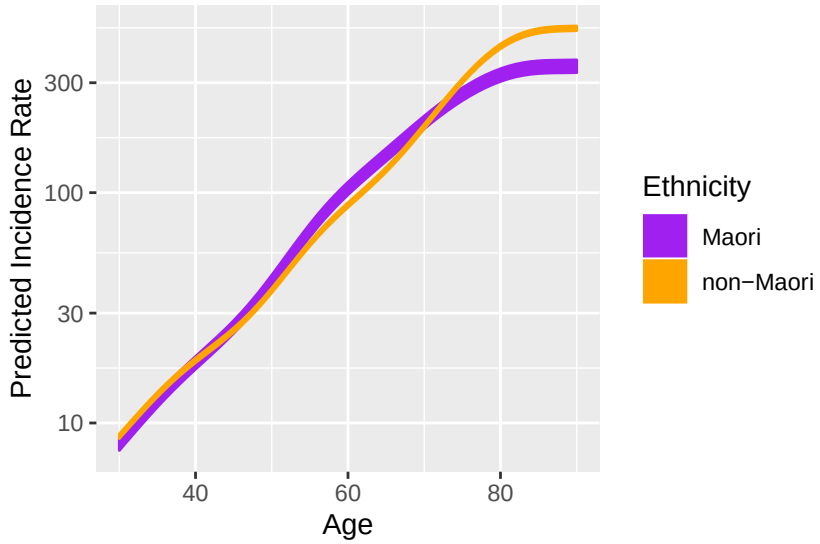


Figure 3.20: *Model-based age-specific CRC incidence rates by ethnicity with 95% CI, estimated by the APC model for 2018 (based on uncorrected counts of CRC cases). The estimates are presented on logarithmic scale in order to make the visual comparison easier.*

3.3.5 Sensitivity analysis

3.3.5.1 Sensitivity to undercount of Māori ethnicity in NZCR

This section includes results from analyses carried out using Lexis cells with corrected counts of Māori and non-Māori CRC cases, constructed by applying the correction factors specified in section 3.2.1.2. As can be seen in Figure 3.21 (for the population 30-<90 years old), the ASRs in Māori were high only in years when the correction of the counts was applied (i.e., before 2006), unlike the flat trend in ASRs based on uncorrected counts (Figure 3.2). The decrease in ASR over time might reflect the correction process, not the actual trends.

The net drift extracted from the APC model fitted to corrected counts of CRC cases was -1.13% (95% CI; -1.58, -0.68) for Māori, and -1.31% (95% CI; -1.42, -1.20) for non-Māori,

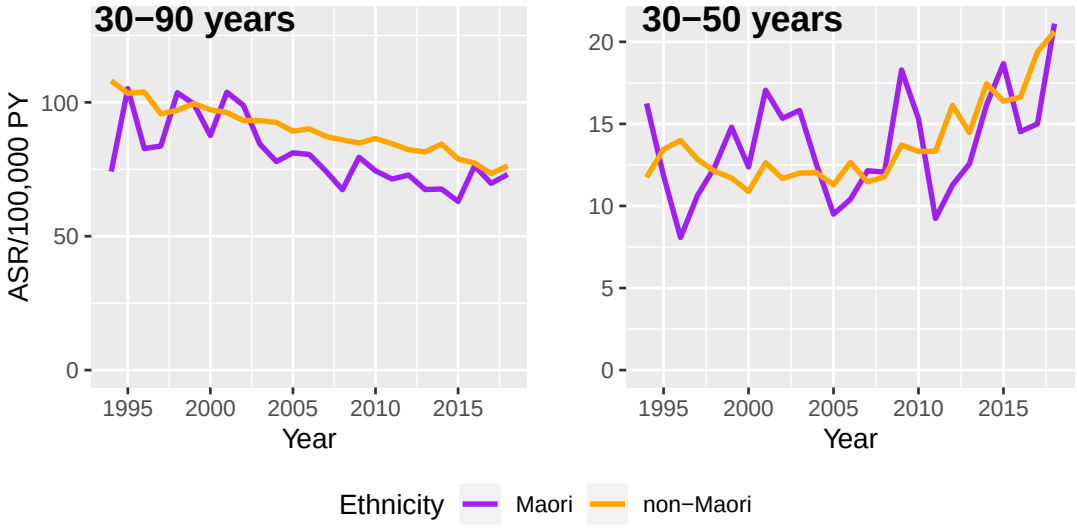


Figure 3.21: *Trends in incidence rates age-standardised to the 2018 NZ population, for age 30-<90 and 30-<50 years for 1-year periods by ethnicity, using corrected counts of CRC cases.*

compared to 0.25% for Māori and -1.38% for non-Māori based on uncorrected counts. The estimated age, cohort and period effects for Māori and non-Māori using data with corrected counts of CRC incidence in both ethnicities are presented in Figure 3.22. The shape of the cohort effect in Māori was different than in non-Māori. When using the corrected counts, successive generations of Māori born from 1904 to 1955 experienced a steady decrease in IRRs. However, the sharp increase in IRRs in generations born from around the 1960s onwards, especially pronounced in generations born between the 1970s and 80s is very similar to the results of the APC model with uncorrected counts (Figure 3.13). Period deviations in both ethnicities are negligible and not clinically relevant (see Section 3.3.3.2).

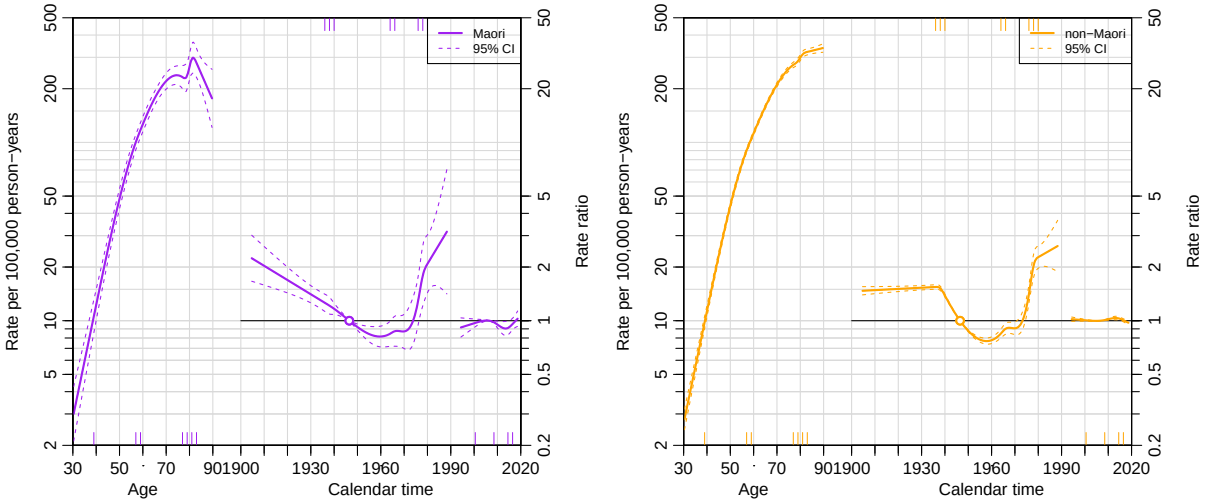


Figure 3.22: *The estimated age, period and cohort effects from the APC model for incident CRC by ethnicity, assuming zero period slope, using corrected counts. The reference is the 1946.5 cohort to which the age-effect refers. The reference period is 2006.5.*

In order to compare the output from both APC models (based on uncorrected and uncorrected counts of CRC cases) the results are plotted in one figure (Figure 3.23). As can be seen, in addition to cohort effects also the longitudinal age effects are different.

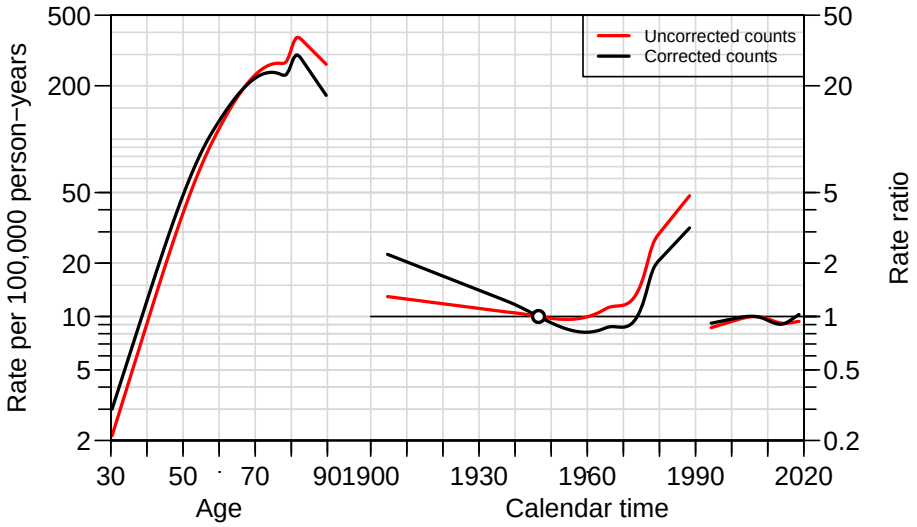


Figure 3.23: *Age, period and cohort effects from models with corrected and uncorrected counts of CRC cases.*

With respect to the cumulative longitudinal age effects based on corrected counts, Figure 3.24 (equivalent to the effect in Figure 3.15 based on uncorrected counts) shows the age at which individuals born in a specific year would be expected to be diagnosed with CRC, assuming that they will be diagnosed with CRC between age 30 and 90 years. Using corrected counts, based on the cumulative longitudinal age effects, the median age at diagnosis would be 75.5 years for non-Māori and 73.0 years for Māori and this difference can be explained by age effects.

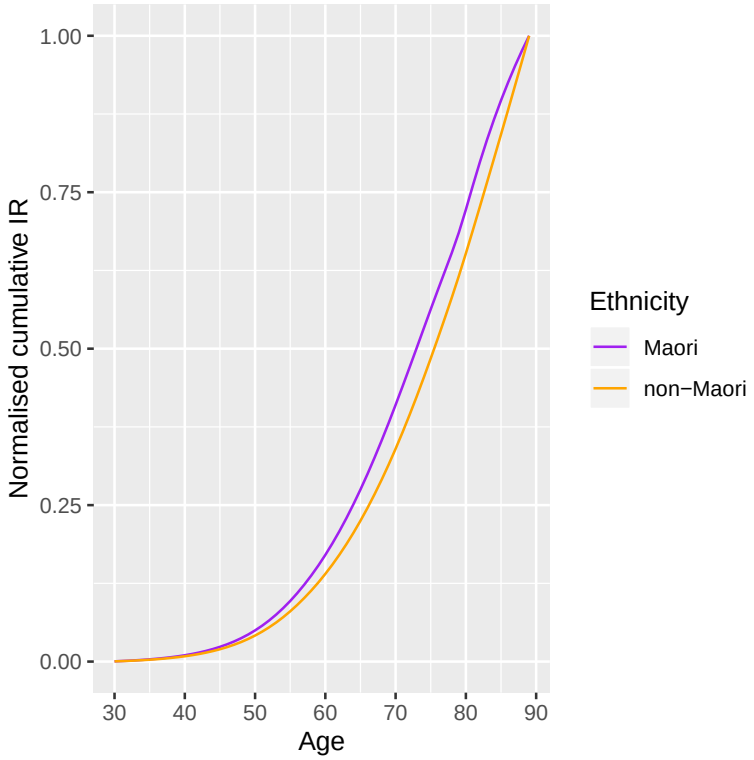


Figure 3.24: *Longitudinal cumulative CRC incidence for registrations of CRC incidence between 1994 and 2018 by ethnicity, for ages 30-<90 years, based on fitted APC models using corrected counts. The cumulative incidence rates (y axis) have been normalised to a cumulative incidence equal to 1 for age 90 years in order to compare the age effects in both ethnicities independently of the overall incidence.*

The graphs for the years 1994, 2002, 2010 and 2018 in Figure 3.25 show how the incidence

rates changed over the study period when the model-based IRs are calculated using data with corrected counts of CRC cases for both ethnicities. Incidence rates based on corrected counts in years 1994 and 2002 are, in general, still lower for Māori than in non-Māori, however the difference is much smaller than that based on uncorrected counts (Figure 3.19).

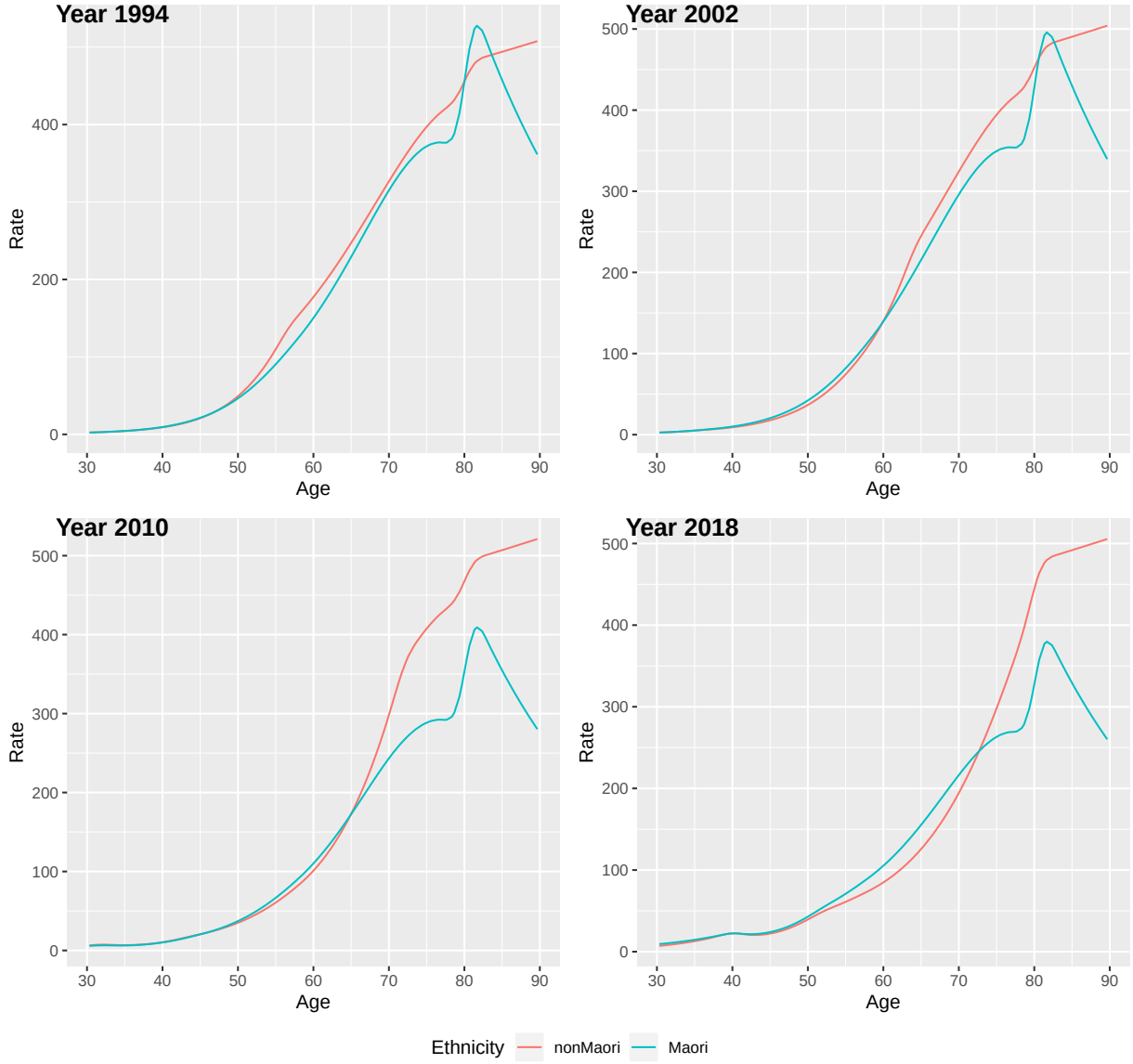


Figure 3.25: *Model-based age-specific CRC incidence rates by ethnicity, using corrected counts, estimated by the APC model for the four chosen periods.*

3.3.5.2 Sensitivity to drift allocation

Figure 3.26 shows the results of the analysis with the cohort slope set to zero and the drift allocated to period compared to the main analysis. Based on this parametrisation the age, period and cohort effects for Māori were very similar as in the main analysis. In non-Māori, the differences are more pronounced. For example, the IRR in generations born in the 1970s and 80s would be even bigger than the values estimated based on zero period slope.

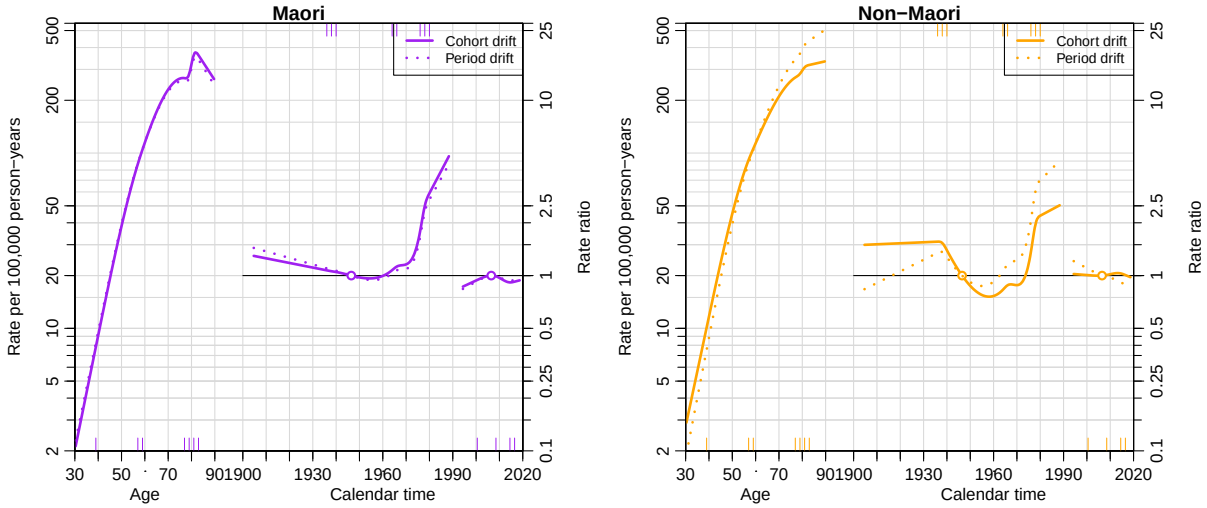


Figure 3.26: *The estimated age, period and cohort effects with drift allocated to period effect by ethnicity. The reference is the 1946.5 cohort to which the age effect refers. The reference period is 2006.5.*

3.4 Discussion

In this section I summarise and discuss the main findings, compare the main results with the literature, explain how predictions from the fitted APC models can be used in clinical practice, discuss the validity of the results, and provide an overview of the strengths and

the identified limitations of this sub-study. The scope for future research that emerged from this sub-study is presented in the overall discussion of the whole PhD study, in Section 6.

3.4.1 Summary and interpretation of main findings

This sub-study described trends in CRC incidence in NZ from 1994 to 2018, which differed between Māori and non-Māori, and identified population strata with especially increased incidence rates. Before I interpret the findings, I would like to remind the reader that the interpretation of the estimates of age, period and cohort effects are valid only under the assumption of zero period slope. The justification of that assumption is given in Section 3.1.2.3. The results are based on a model which relies on the strong assumption that the age-period interactions are explained by a cohort effect. There was no indication of disagreement between the modelled and observed rates in my analysis, which gives confidence that the cohort effect, not age-period interactions, explains the patterns in the data. I am aware that failure to reject the null hypothesis does not prove that the model actually is correct. However, given the size of the analysed data set I would expect any substantial deviations from the model's predictions to be picked up by the chi-square test. The fact that Carstensen's method for knot selection for the splines led to rejection of the hypothesis for the model fit supports the idea that the chi-square test has a reasonable assay sensitivity for this data set. Additionally, the initial inspection of the four classical plots indicated that the data do not follow an age-period model, a crucial explorative observation which was be made before imposing strong modelling assumptions. Finally, another model assumption, namely that age, period and cohort effects are smooth and therefore can be modelled using splines, was supported by the goodness-of-fit analysis.

This sub-study has three main findings. First, the analysis shows that in the NZ popu-

lation born in the 1970s and 80s, regardless of gender and ethnicity, the CRC incidence rates ratios increased sharply with later year of birth. Based on analysis of incidence data from 1994 to 2018, the IRRs in generations born in the 1980s were the highest estimated among all birth cohorts, and over three times as high as for those born in the 1960s. This is an alarming finding, because the model suggests that when the older generations (with lower IRRs) are replaced by the younger generations (with highly increased IRRs) a high wave of CRC diagnoses can be expected in NZ.

Cohort effect has this alarming consequence because unlike a period effect, which can change at any time due to e.g. a change in disease classification or incorporation of better diagnostic procedures, a birth cohort effect, per definition, does not change over time. An individual born in 1970 will always have the birth cohort effect associated with birth year 1970 and, therefore, the increased or decreased IRR due to the cohort effect will follow each birth cohort throughout life. However, since future period effects are not known, the future incidence rates for a specific birth cohort can change if a period effect changes substantially. A decrease in incidence rates due to a change in period effect could come from e.g. changes to diagnostic procedures that would allow more polyps to be found and removed before they become cancerous.

Second, the analysis of the overall CRC incidence data showed a sharp decrease in IRR in successive generations born from around 1940 to around 1955, preceded by a flat trend in generations born between 1904 and 1939. The decreasing cohort effect can explain the decreasing ASRs in the NZ population between 1994 and 2018, despite the lack of population based CRC screening which is widely claimed as one of the reasons for declining IRs in other countries.

Third, the fact that the APC model fitted the data well and that the period deviations were negligible suggests that trends in CRC incidence are driven primarily by generation-specific environmental exposures that take place mostly in childhood and early adulthood

(Murphy and Yang, 2018) and influence CRC risk during the whole life, independently of age and period of diagnosis.

Related to the third finding is a controversy about whether early onset of CRC (EO-CRC), i.e. CRC in individuals younger than 50 years, is distinct from late onset CRC (LO-CRC), i.e. CRC in patients older than 50 years. While some studies provide morphological and genetic evidence for a distinction between EO-CRR and LO-CRc (Cavestro et al., 2018; Hofseth et al., 2020), the results of the recent study by Dharwadkar et al. (2019) (in the US population) do not support that distinction. The epidemiological evidence from this APC analysis, in combination with the findings of Cox and Little (1992), does not support the distinction in the NZ population either. The decrease in incidence rates in young age groups identified by Cox and Little (1992) and attributed to birth cohort effect (1937–1957) was mirrored by a decrease in incidence rates in older individuals in the same birth cohorts identified in this sub-study. This shows that in NZ the EO-CRC and LO-CRC are related to the same birth cohorts. It can therefore be suggested that the late and early onset of CRC are also related to the same environmental exposures, which in turn suggests that they should be seen as one disease.

Additionally, the results of the analysis by sub-sites are interesting and worthy of mention. Although the cohort effects were similar in all three sub-sites in males and in females, the cohort effects in generations born in the 1980s showed a decreasing trend in distal colon cancers in males and females. This is an optimistic message, but has to be treated with caution due to a very small number of tumours in those born after 1980, resulting in very wide 95% CIs.

The study aimed also at describing trends separately for Māori and non-Māori. Based on this study it is not possible to provide an unambiguous description of the trend in Māori, as the cohort effect prior to 1955 is very sensitive to the correction of the undercount of Māori in NZCR before 2006, as shown by the sensitivity analysis (Section 3.3.5.1).

However, several results about trends by ethnicity can be provided as they are not sensitive to the correction of the counts. First, the analysis stratified by ethnicity revealed that the very sharp decrease in IRRs in successive generations born from 1940 to 1955 affected only non-Māori, while in Māori the IRRs declined linearly in generations born between 1904 and 1955. This finding may help towards a better understanding of the historical changes in CRC incidence in those two ethnicities. Second, the fast growing trend in generations born in the 1970s and 80s which affected equally both ethnicities is only minimally affected by the undercount because individuals born in the 70s and 80s are unlikely to get CRC earlier than in 2006 due to their young age. Also, as the predicted CRC incidence rates in Māori after 2006 are only slightly affected by the correction for the undercount, the finding that in 2018, Māori aged from around 50 to 70 years have higher IRs than non-Māori is robust.

Finally, one of the objectives of this sub-study was to explain the lower median age at diagnosis in Māori than in non-Māori frequently mentioned in the literature. In the analysed data the difference was 7.5 years and most of the difference (5.5 years) can be explained by the population structure. The remainder which is around two years can be explained nearly entirely by the cohort effect (for uncorrected counts), or nearly entirely by the age effect (for corrected counts).

3.4.2 Comparison with earlier studies

This section begins with a comparison of this sub-study's results to the trends in CRC incidence in the NZ population reported by other researchers. I then go on to compare the cohort effects found for NZ in this sub-study to those reported in different populations.

The steep decrease in IRRs in birth cohorts around 1939–1955 was already reported in a NZ study by Cox and Little (1992), who analysed CRC incidence data, from the years

1957–1986. Cox and Little (1992) did not perform APC analysis but presented incidence data using 5-year birth cohorts by 5-year age groups (the same presentation I showed in Figure 3.4 D). They found a decrease in incidence rates in the successive birth cohorts 1937–1957 who were at the time of the study young individuals (25–39 years old). Cox and Little attributed the decrease to the cohort effect, but without carrying out an APC analysis it was not possible to test whether the decrease was indeed explained by cohort effect. Their study showed a similar pattern to my results, but in Cox and Little's study those generations were young while in this sub-study those generations were between 50-<75 years old. The decreasing incidence with increasing year of birth in those generations (born between 1939 and 1955), is reflected in the current decrease in age-standardised incidence rates in NZ. The low incidence rates have been following those generations during their whole life - exactly what the cohort effect in the APC model means.

The age-specific trends in CRC incidence rates found in this sub-study are generally consistent with the findings reported by others who included NZ data in their analyses. A recent study by Wong et al. (2020) analysed time trends in age-standardised CRC incidence rates using join-point regression covering the period from 1983 to 2012 for NZ. Wong et al. (2020) showed declining trends in CRC incidence in people 50 years and older but increasing incidence in the NZ population younger than 50 years for colon and for rectal cancers. The estimates in the young individuals, however, had wide CIs. In this sub-study, I analysed data covering a period of diagnosis from 1994 to 2018, and the increasing incidence rates in young age groups were clearly statistically significant, as shown by the 95% CIs for the local drifts (Figure 3.16). The following two factors can explain the discrepancy: firstly, I analysed six additional years, which gave me more data for the late birth cohorts affected with highly increased incidence; and secondly, the APC analysis allowed me to analyse the data more efficiently than join-point regression, which does not borrow information between age groups.

The results of my analysis showing decreasing incidence unique to ages 50+ but increasing

incidence in generations of young adults born from the 1970s onwards were also broadly consistent with the findings of Siegel et al. (2019). There are small discrepancies, e.g. the AAPC of 4% reported for 20–49 years old between years 2007–2016 in Siegel was of similar magnitude in my analysis but applied only to age 30-<40 years (Figure 3.9). In this sub-study it was possible to find the exact age to which the increase applied due to the use of 1 year age brackets, unlike Siegel, whose use of very broad age brackets prevented more precise results.

The comparison of this sub-study's results to the results of the only published APC analysis of CRC incidence data from NZ by Araghi et al. (2019) in general shows a broad consistency. However, the IRRs in the youngest birth cohorts analysed in this sub-study are lower than those reported by Araghi et al. (2019). Araghi et al. found five-fold increases in late birth cohorts (compared to the birth cohort with the lowest IRR), whereas in this sub-study the estimated increase was just three-fold higher. Araghi et al. also aimed at comparison between trends in seven high-income countries and therefore, reasonably, did not concentrate on specifically fitting NZ data (there are no results in Araghi et al. concerning the goodness of fit analysis with respect to NZ data).

In this sub-study, by contrast, the APC model was specifically optimized to fit the NZ incidence data, and the results presented in this sub-study would therefore be expected to be more precise. Additionally, the age brackets used in the analysis for presenting trends and the AAPC in Araghi et al. (2019) are very broad (0–49 years, 50–75 years and 75+ years). In the analysis of trends for the population younger than 50 years, in Araghi et al., the age brackets were also very broad (10-year age brackets: 20–29, 30–39, 40–49) and included only data from the period 2005–2014. Therefore, in addition to the more precise estimates, this sub-study shows the whole picture of trends in CRC incidence in NZ for the period 1994–2018 for 1-year age groups. Additionally, this sub-study includes analysis by ethnicity and analysis for three anatomical sub-sites by gender, which is an important addition to findings that have already been reported when considering epidemiology of

CRC in NZ.

A direct comparison of my results to one of the NZ studies which investigated temporal trends in CRC incidence, Shah et al. (2012), is limited, as Shah et al. analysed data from NZCR covering much earlier periods (1981–2004). Nevertheless, the different values of net drifts in females in proximal cancer and distal cancers (Table 3.6) found in this sub-study confirms the left to right shift in females reported previously by Shah et al. (2012). What this sub-study adds is an explanation that the left to right shift is partially due to the cohort effect and only partially due to the ageing of the NZ population. Thus, the increasing proportions of proximal tumours in females is due to the increasing IRRs for proximal cancers with increasing year of birth in females born between 1904 and 1939, while for the same birth cohorts IRRs for distal cancers were flat (as shown in Figure 3.17). The shift is also, but only partially, due to age effect, as proximal cancers are more common in older females (Figure 3.17), which in combination with the ageing of the NZ population results in more proximal cancers compared to distal cancers over time in females. In this analysis there was no evidence for left to right shift in NZ males. This finding is similar to the results described by Shah et al. (2012), but contrasts with findings from an APC analysis conducted in Norway (Larsen and Bray, 2010), where the left to right shift was observed in both genders and associated with birth cohorts 1900–1950.

The results of the analysis for anatomical sub-sites based on APC models conducted in this sub-study differ from those reported in a NZ study by Gandhi et al. (2017). Gandhi et al. (2017) found increasing incidence in young New Zealanders, especially in males, limited only to rectal tumours, whereas this sub-study found increasing incidence rates in all three sub-sites and in both genders. The analysis conducted in this sub-study is probably more accurate than that of Gandhi et al., due to the modelling of age and calendar time on continuous scales and the inclusion of newer data. Moreover, Gandhi et al. (2017) did not investigate whether cohort effect can explain the increase in IRs in

young New Zealanders, and most likely attributed his findings to age-period interactions. By contrast, this sub-study concluded that cohort effect is responsible for the increase in IRs in young New Zealanders.

The trends in CRC incidence described in this sub-study are not NZ specific. Similar trends have been observed in various countries with a high Human Development Index, including Australia, the US, Canada and several Western European countries. These countries show declining or stable trends in colon and rectal cancer incidence in people aged >50 years but increasing incidence among populations aged <50 years (Arnold et al., 2017; Wong et al., 2020).

The strong cohort effects identified in this sub-study as a part of the trends in CRC incidence are also not specific to the NZ data. In studies which used APC modelling for the analysis of CRC incidence, substantial cohort effects responsible for the observed incidence trends were consistently identified. However, cohort effects differed slightly between countries with respect to which birth cohorts were affected, which cohorts experienced decrease or increase in CRC incidence, and the size of the effect. The analysis of incidence data from Norway by Larsen and Bray (2010), who analysed the period 1962–2006, showed an increasing trend in the whole population in the first two decades but stabilisation of the rates for colon cancers only in younger ages in periods since the 1980s. The authors expressed concern about increasing rates of rectal tumours in the younger age groups during the studied period. In addition to the much more pronounced cohort effect, the data from Norway showed a pronounced period effect, which the authors attributed partially to the increased intake of screening in the early 80s in the Norwegian population. Studies conducted in other Nordic countries also found and reported effects of birth cohort on CRC incidence rates (Dubrow et al., 1994; Thörn et al., 1998; Svensson et al., 2005).

The highly increased incidence rate ratios in generations of New Zealanders born in the

1970s to 90s was similar to that reported in a more recent APC analysis of CRC incidence data from Canada by Brenner et al. (2017), who found a dramatic cohort effect resulting in increased IRRs in young generations of Canadians. However, the strong decrease in IRRs in generations born between 1940 and 1955 in Canada was observed only for colon cancer, while in NZ the decrease related to all three anatomical sub-sites.

The increasing IRRs in younger generations found in this sub-study is similar to the temporal CRC trends in the Australian population reported in Feletto et al. (2019), who also found generations born between the 1960s and 90s to be affected by increased incidence rates for both colon and rectal cancers. A similar association was reported in the US population (Siegel et al., 2017; Sung et al., 2019). The cohort deviations in generations born from the late 1960s to 80s identified in this sub-study are very similar to the cohort deviations presented for Australians (Feletto et al., 2019) and Canadians (Brenner et al., 2017). However, the cohort deviations in the NZ population (strictly, in the non-Māori population) who experienced a very sharp decrease in IRRs in generations born between 1940 and circa 1955 are slightly stronger than those found in Feletto et al. (2019) and Brenner et al. (2017).

A different pattern was found in the Nordic countries, where the IRRs generally increased for generations born between 1870 and 1950. The exceptions to this pattern were Norway and Estonia, where the decrease in CRC incidence was observed in generations born during World War II (Svensson et al., 2005). The deep decrease in IRRs in generations born during World War II in Norway and Estonia can be attributed to calorie restriction during early life. This explanation does not apply to the NZ non-Māori population, as most likely those born between 1940 and 1955 in NZ did not experience drastic decreased calorie intake during early life.

A recent paper by Chung et al. (2019) used an APC model to analyse CRC incidence data from Western countries (the UK, the US and Australia) and from Asia (Japan, Hong

Kong, Shanghai, Singapore and India), but included the incidence up to year 2007 only. The paper reported a cohort effect in young populations in all these countries, with very similar patterns to those found in the NZ population by this sub-study. The authors expressed concerns about an increasing risk of CRC in all these countries when their younger cohorts reach older ages; this is the implication of the pronounced cohort effects. Based on the APC analysis presented in this sub-study, the same concern applies to the NZ population. Indeed, the worry is that the cohort effect, which represents the exposure to environmental risk factors during early years of life, will follow the affected generations over their life course. Possible solutions to this alarming problem include alteration of the lifestyle-related exposures that cause increasing incidence with increasing year of birth. However, these exposures can differ between countries, and it is necessary to generate and test hypotheses in future studies to identify the specific relevant exposures in NZ. This sub-study's results can be helpful in generating such hypotheses.

3.4.3 Use of the results in clinical practice

Although the application of the results in clinical practice is not the main focus of this study, in this section I discuss some ideas about how the results can be communicated to health professionals, and how policy-makers and clinicians could make use of the results.

As previously explained in Section 3.2.2.2, the estimated age, period and cohort effects depend on the parametrisation, and those values will change with different slope allocation and/or with a different choice of reference groups. The model-based incidence rates do not depend on parametrisation and as such are more suitable for communicating to health professionals than the estimates of age, period and cohort effects. However, as the predicted incidence rates depend on the model, the choice of models which provided appropriate fit to the data was important.

Although the evidence communicated to GPs should, ideally, be based on the primary care population data, in which case the denominator would be the population registered with GPs, the IRs in the primary care population would differ only slightly from the IRs in the general population because nearly all New Zealanders are registered in primary care. Due to the lack of a primary care research database in NZ, incidence rates in primary care can not be estimated in NZ, and therefore this sub-study proposed incidence rates for the primary care practice in NZ based on data from the general population [2].

The predicted IRs in population strata communicated to doctors, e.g. as the plots shown in Figure 3.18, could be useful during assessment for further investigation of any patient 30-<90 years old. The proposed plots allow the quick initial assessment of the baseline risk for each of the anatomical sub-sites (for year 2018, stratified by age and gender). This knowledge, combined with patients' symptoms, test results and comorbidities, could help to select patients for a specific type of further investigation. Patients with increased risk of rectal and distal tumours could initially be offered less invasive and more available investigation, e.g. digital examination or flexible sigmoidoscopy, which could help in better management of the existing capacity of colonoscopies [3]. The incidence rates estimated for the year 2018 based on the APC model can also be used by policy makers, in developing new policies related to CRC or updating the already existing policies.

[2] I raised this issue on Research Gate and Stack Exchange. The consensus was that incidence rates based on data collected from the general population are suitable for use in primary care; however, these incidence rates will be slightly underestimated relative to those based on primary care data.

[3] In private correspondence, Marcela Ewing, a Swedish medical doctor who conducted PhD research into early diagnosis of cancer, wrote: *"There is no way the GP can estimate from what part of the intestine the symptoms come from as it is so hard already to diagnose CRC unless you send the patient for colonoscopy. So unfortunately, I don't think that your prediction model can be used clinically yet, or not in the present stage."* Because, according to Ewing, symptoms alone can not help GPs to decide in which part of the colon the tumour is located, it follows, in my opinion, that the model proposed in this sub-study would be useful, since the model gives probabilities of the tumours being located in the proximal or distal colon.

Below I explain why these results are more suitable for use in clinical practice and policy-making than estimates based on simpler approaches. Compared to simpler approaches such as age-specific and age-standardised rates, incidence rates calculated using the APC model have an important advantage for use in clinical practice. Before the results of studies can lead to implementation in policies and guidelines, the data will typically be several years old. It is therefore important that the research used by policy makers gives an accurate description of the situation at the end of the study, which is difficult to achieve without APC analysis. This is because the estimates of IRs from APC models for the last year of the analysed data are based on the whole data set, from 1994 to 2018. The crude age-specific incidence rates for the last study year will be based on too small sample size to give narrow CIs, unless reported for very broad age groups. The use of broad age categories is not an optimal solution, as CRC rates are not the same in e.g. 50- and 59-year-old people (which is assumed when 10-year brackets are used). APC models ensure narrow CIs (Rosenberg and Anderson, 2011); these are vital for decision making in clinical practice and imply a smaller risk that policy based on point estimates could be wrong.

An example of a proposed policy that could be informed by the results of the APC analysis is the suggestion made by the MoH (Ministry of Health NZ, 2018b) to lower the screening age for Māori by 10 years. The rationale for this suggestion is that Māori currently benefit less from screening than non-Māori in terms of life years gained, and widening the screening age for Māori would help to achieve equity. Alternatively, the decision could be based on another criterion, that is, according to the risk of CRC. Based on my results it could be suggested that screening of Māori should start about two and a half years earlier (at age 57.5 years) to maximise the benefit from the screening program for the whole NZ population (see Graph 3.19 for the year 2018 [4]).

[4]The curves are based on analysis which used data with uncorrected counts, however the curves for the year 2018 based on corrected counts give very similar result.

It has already been noted in earlier NZ research (Gandhi et al., 2017) that knowledge about the increased rates of CRC in young adults can help doctors to make decisions in clinical practice. Knowledge about increased rates of CRC can inform decisions about surveillance in young adults who present with symptoms that could indicate CRC, as well as decisions about prompt investigation, including colonoscopy. The results from this sub-study present a sharper picture of the incidence rates in young adults. In particular, this sub-study indicates that the increase in incidence affects entire generations, not only young adults; moreover, the results of my analysis show that the increased incidence applies to all three anatomical sub-sites in both genders. Without using the APC model it would seem that only young individuals are affected. However, the apparent association with young age is due to the current data availability for recent birth cohorts being restricted to patients younger than 50 years.

3.4.4 Validity of the results

The internal validity of the sub-study was satisfactory. There was no disagreement between fitted and observed rates, as was demonstrated in the goodness-of-fit analysis. Since this was a population-based study there are no issues with the representativeness of the sample. Due to the legal requirement that from 1994 registrations of cancers in NZCR are mandatory, the data can be considered complete, so the sub-study is unlikely to be affected by reporting bias. The quality of NZCR data has been reported to be good, with the exception of the staging information, which was not used in fitted models.

The external validity of sub-study 1 is not a matter of concern, as the results are not meant to applied outside of the NZ population. However, care has to be taken when considering the implications of the results for clinical practice. The studied population was the general NZ population, and therefore the IRs differ from those estimated from data collected from primary and secondary care.

3.4.5 Strengths

One of the strengths of this sub-study is the use of the APC model for analysis of CRC incidence data, which allowed: separation of the cohort effect from the age and period effect; quantification of the increasing CRC rates in younger populations of New Zealanders; and, importantly, the identification of the generations that experienced the increased rates.

The data were analysed using one-year resolution for all three time-scales. This sub-study's estimates of the effects therefore have higher precision than studies which used more coarse tabulation (e.g. 5-year age or period brackets). An additional advantage of the fine tabulation is that the identified cohort effect was appropriately adjusted for age and period.

An additional strength of this sub-study was the choice of the methods used to present the time trends. I used a wide range of presentation methods to explain both the incidence rates and the complex results from the fitted APC models. I also provided examples of model-based curves for the direct comparison of incidence rates that could be used in clinical practice.

Because the estimates of the effects are based on the critical assumption about the lack of interactions between the time scales, I carried out a careful assessment of the violation of the modelling assumption. This gives confidence that the results of the analyses are robust and are suitable for proposal for use by NZ health professionals.

3.4.6 Limitations

The sub-study has some limitations which have to be acknowledged. First, the age, period and cohort effects were estimated under the assumption of zero period slope which

might not be an appropriate choice for the analysis of CRC incidence in Māori as the undercount before 2006 creates a period effect which in the analysis is attributed to the cohort effect.

The allocation of the drift to the cohort effect is an arbitrary choice, but this choice was informed by the knowledge that there were, except for the undercount of Māori, no major events during the period studied, such as a change in classification of CRC in NZCR or the introduction of a screening program. Therefore the assumption of zero period slope for was probably appropriate, except for the analysis for Māori. Further, the allocation of drift to cohort has been used by earlier researchers (e.g. Larsen and Bray (2010); Araghi et al. (2019)). In this analysis the choice was informed by the evidence of very pronounced cohort deviations, while the identified period deviations were trivial.

Second, with respect to subgroup analyses, the estimates are based only on separate analyses of individual strata, as the Epi package does not provide an implementation of the model that could use the whole data set and include covariates. My programming skills are not good enough to implement such a model by myself. However as reported by Rutherford et al. (2010) the difference in estimates from those two approaches would be rather small.

Third, changes in ethnic self-identification could potentially influence the trends in incidence rates in both ethnic groups. For example, the similar incidence rates in younger generations could be influenced by the increase over time in the proportion of Māori who are of mixed ethnicity.

Finally, it is debatable if the selection of the knots for the splines based on backward elimination is appropriate. According to Simpson (2018b), backward elimination might not be the best choice as the different models corresponding to different sets of knots are not nested. However, I have chosen to use backward elimination for the choice of the knots because, as explained by Dobson and Barnett (2008), BIC and AIC are especially suited

for comparing non-nested models. Also, there was no evidence for lack of fit to the data for any of the fitted models which gives confidence in the appropriateness of the method used for knot selection. In my opinion, therefore, this choice is not a limitation.

3.5 Conclusions

The analysis of CRC incidence data using APC analysis provided an accurate and illuminating picture of CRC incidence trends in NZ between 1994 and 2018. The analysis found that the increase in CRC incidence in young New Zealanders in recent years is almost entirely due to the cohort effect. This is an alarming finding, as the cohort effect will follow the affected generations throughout their life. Based on this sub-study's findings, a wave of CRC diagnoses can be expected in the future in NZ, when younger generations with high incidence of CRC replace older generations with lower incidence of CRC. This sub-study provides a very detailed picture of which generations have low incidence of CRC and which generations are affected by the greatly increased incidence of CRC; this knowledge can be used for planning the allocation of the future resources needed to deal with the expected wave of new CRC diagnoses.

To the best of my knowledge, this is the first study which has disentangled the effects of age, birth cohort and period of diagnosis separately for Māori and non-Māori, however this study was not able to resolve the problem with undercount of Māori and therefore to provide reliable estimates of the contribution of age, cohort and period effects to the incidence rates of CRC in Māori.

This sub-study found no disagreement between the predicted and observed data, which gives confidence that the results are reliable, except for the results of APC analysis of incidence rates in Māori before 2006. The predictions of IRs from the fitted APC models (overall, for males and females, for proximal, distal and rectal cancers by gender, and by

ethnicity) can therefore be proposed for use in NZ clinical practice and policy-making. Also the predicted IRs for Māori and non-Māori in the year 2018 are reliable as they do not depend on the use of corrected or uncorrected counts and can be a valuable source of information for clinical practice and policy-making.

Chapter 4

Association between diabetes and colorectal cancer in NZ patients with relation to diabetes duration and insulin use

4.1 Introduction

This chapter presents sub-study 2, which investigates diabetes as a risk factor for CRC in the NZ population, in addition to the risk factors which were explored in sub-study 1[1]. The next part of this section provides a brief overview of the evidence for diabetes-CRC associations including discussion of factors that may be important in analysis to determine the association. More detailed background information about diabetes and its link with CRC was provided in Section 2.5.4.5 of Chapter 2. Subsequently, the rationale for the research, aim and objectives for sub-study 2 are specified.

4.1.1 Association between diabetes and CRC

The association between diabetes and CRC is well established (Tsilidis et al., 2015). Several meta-analyses of cohort and case-control studies reported the pooled estimated increase of CRC risk in persons with diabetes compared to non-diabetics or to the general population to be around 30% (Larsson et al., 2005; Jiang et al., 2011; Starup-Linde et al., 2013; Wu et al., 2013; Luo et al., 2016; Sacerdote and Ricceri, 2018). There are, however, differences in the strength of the association estimated in different studies. The differences are captured by the heterogeneity among studies reported in almost all meta-analyses, except for the meta-analysis by Larsson et al. (2005). In addition, (González et al., 2017) showed that prevalence of diabetes does not correlate with CRC incidence on a worldwide basis. Such findings suggest likely differences in the association between diabetes and CRC in different populations.

There are a range of possible explanations for the differences in the effect of diabetes

[1]The preliminary results of this study were presented: as a poster in the Annual Meeting for the NZ Diabetes Association 2019 in Napier; as a presentation at the Postgraduate Conference at the University of Waikato in 2019; and as the oral presentation for patients with diabetes from Hamilton during one of the monthly meetings in 2019.

on CRC incidence in different populations. The differences might be real, e.g., due to the different exposure to factors associated with both diseases, such as environmental exposures, physical activity level, lower abdominal obesity and diet. Alternatively, the differences could be due to confounding with common risk factors. Finally, the differences might be due to biases in the studies' design not addressed appropriately in the analyses. One of the most important biases is detection bias, which can cause the overall estimate to be falsely increased. This bias is a particular risk in studies which have a short follow-up and include only incident diabetes in the statistical analyses. Studies which included duration of diabetes, and thus investigated detection bias, showed that incidence of CRC is especially increased shortly after diabetes diagnosis, which suggests the higher CRC incidence rate may result from the increased medical surveillance in patients with newly diagnosed diabetes (Carstensen et al., 2012; Johnson et al., 2012). The association between diabetes and CRC can also be modified by use of anti-diabetic medication, with insulin being of a particular interest because many studies demonstrate increased CRC incidence in insulin users compared to non-insulin users (Yin et al., 2014).

Despite the modest magnitude of the association, diabetes status has been proposed as an additional factor that doctors can apply in clinical practice when investigating patients for presence/absence of CRC (Starup-Linde et al., 2013). It is not obvious however, that such proposals are valid for NZ practice, as such associations still need to be confirmed using data from the NZ population. A recent study by Mikaeel et al. (2021) found association between diabetes and CRC to be especially increased in young patients (18–54 years old), and also proposed diabetes status as a factor for better selection of patients for further investigation. Limburg et al. (2005) have also proposed diabetes status for possible stratification of the population for CRC screening. In particular, Limburg proposed screening of postmenopausal females with diabetes earlier due to the increased CRC risk in this strata. The usefulness of such approaches will depend on the strength of the association in the local populations.

The use of diabetes status in clinical practice for prognostic purposes could be especially important in NZ, a country with high incidence of CRC and a moderate to high prevalence of diabetes. Additionally, diabetes prevalence in the Indigenous part of the NZ population (Māori) has been shown to be twice as high as in non-Māori (Atlantis et al., 2017), while Māori also have poorer outcomes after CRC diagnosis (Hill et al., 2010; Sharples et al., 2018). This context makes it especially important to estimate the association between diabetes status and CRC incidence for Māori.

The next two sections discuss confounders and effect modifiers that are especially important in studies investigating associations between diabetes and CRC.

4.1.2 Confounders in association between diabetes and CRC

Confounding with common risk factors can be one of the factors responsible for the reported associations between diabetes and CRC (Tsilidis et al., 2015). Therefore, in studies investigating causal effects of diabetes on CRC risk, failing to adequately control for potential confounders can lead to misleading or inconclusive results. In order to demonstrate a causal effect of diabetes on CRC risk, the analysis therefore needs to control for known common risk factors such as obesity or physical activity level (González et al., 2017). However, in studies which aim at using the strength of the associations for diagnostic purposes, the lack of controlling for many potential confounders is not a problem, as long as the statistical model includes the main predictors known to policy-makers and to medical professionals, that is: age, gender and ethnicity. This is because, in studies investigating the association between diabetes status and CRC for diagnostic purposes, diabetes status would act as a surrogate for unobserved predictors (e.g., obesity or lack of physical activity). For clinicians and policy-makers it is not important whether the increased incidence of CRC in diabetes is an effect of diabetes itself or an

effect of confounders. Therefore, despite the lack of evidence for causal effects, robust evidence for the increased incidence rates of CRC in patients with diabetes (in earlier studies around 30%) has been already suggested for use in clinical practice as diagnostic factor (Giouleme et al., 2011; Starup-Linde et al., 2013).

4.1.3 Effect modifiers of association between diabetes and CRC

In studies on the association between diabetes and CRC, the term *effect modifier* is used in two different ways. Firstly it is used interchangeably with *interaction effect* (VanderWeele, 2009)[2] and, in this case, it is a third variable which modifies the diabetes-CRC association and relates to exposed and unexposed individuals. If, e.g., the third variable is gender, the IRRs can be calculated separately in males and in females and, due to the independence of the two samples, the confidence interval for the difference can be calculated by pooling standard errors. Secondly, the term *effect modifier* is also used to describe factors that only apply to exposed individuals. An example is duration of diabetes. In that case, the denominator will be the same for each of the brackets of the duration, and the calculated IRRs will not be based on independent samples. In the following discussion, both types of factors will be referred to as *effect modifiers*.

In diabetes-CRC association studies there are two important factors that can influence the risk of CRC solely in the exposed part of the study population (in this case patients with diabetes): duration of diabetes; and the use of medication to control hyperglycaemia. As reported in most earlier studies, duration of diabetes modulates the CRC risk following a very specific pattern, characterised by a highly increased IRRs in the first few months

[2] VanderWeele (2009) explains that actually, in studies investigating causal effects, there is a subtle difference between those two concepts, which however is not a concern in studies which do not aim at showing a causal effects.

after diabetes diagnosis followed by a decrease in IRRs in the first year, which, in most studies, levelled up around the second year after diabetes diagnosis and remained slightly increased for several years (Johnson et al., 2011; Carstensen et al., 2012; Harding et al., 2015; Dankner et al., 2016). For example, Johnson et al. (2011) analysed data from British Columbia including 370,000 patients with the median follow-up just over 4 years, and reported HR equal to 2.89 (95%CI; 2.22, 3.75) just after diabetes onset. In the later period, 3 months to 10 years, the risk of CRC remained elevated, with HR=1.15 (95%CI; 1.05, 1.25) (both HR adjusted for age, sex, year of diagnosis and the number of physician visits). Carstensen et al. (2012) reported a similarly high risk shortly after diabetes diagnosis in the Danish population, which decreased during the following year to a moderate level, confirming the pattern found by Johnson et al. (2011). This high incidence of CRC just after diabetes diagnosis suggests the presence of detection bias, which is a result of the increased medical surveillance of patients after diabetes diagnosis and, therefore, often leads to the early detection of already-present, but not yet diagnosed, malignancy (Johnson et al., 2012).

The use of medication for controlling hyperglycemia has been also found to modify the CRC risk in patients with diabetes. In some studies, metformin has been shown to reduce the risk of CRC (Suissa and Azoulay, 2012), while exogenous insulin, in general, has been shown to increase the risk of CRC in diabetic patients (Yin et al., 2014). With respect to insulin, the concern relates mostly to the use of analogues to human insulin such as insulin glargine and detemir, which, in some studies, were shown to increase the risk of CRC. The evidence is inconsistent, however. In 2009, four papers were published which presented evidence that the use of insulin analogues has an especially strong effect on increased risk of some cancers, among which was CRC (Colhoun et al., 2009; Currie et al., 2009; Hemkens et al., 2009; Jonasson et al., 2009). Those studies, however, were heavily criticised for presenting biased results (Pocock and Smeeth, 2009). A recently conducted study by But et al. (2017), based on a large data set including combined data from five

countries, addressed the most crucial biases affecting studies on associations of medication use and cancer risk, and found no differences in the risk of CRC in users of insulin analogues glargine and determir compared to the risk in users of human insulin.

The well-established evidence for the association between diabetes and CRC in combination with the high prevalence of diabetes in NZ, led to the rationale, aims and objectives presented below.

4.1.4 Rationale

Investigation of the association between CRC and potential risk factors is of interest for NZ epidemiological research as the results can help address the high burden of CRC. The high prevalence of diabetes in NZ makes diabetes a risk factor of particular interest, particularly given that research from other countries has established an association between diabetes and CRC (Section 2.5.4.5). On this basis, it has already been proposed that doctors should be attentive to patients with diabetes when assessing the need to investigate patients for presence/absence of CRC. However, because the strength of the association differs between populations, it is not obvious whether NZ doctors should also be attentive to diabetes status. Additionally, if the strength of the association reported in other countries holds true for the NZ population, diabetes status could be used for stratification of the population for CRC screening, as previously proposed for the US population (Limburg et al., 2005). This investigation can be important especially for Māori, due to the high prevalence of diabetes in that population, and could therefore be an important step in achieving equity in CRC outcomes in the NZ population. To my best knowledge, however, the association between diabetes and CRC has not yet been studied in NZ, despite the availability of good quality population-based data related to diabetes and cancer, in particular VDR and NZCR.

4.1.5 Study aims and objectives

Motivated by the problems with timely diagnosis of CRC in NZ, as well as by the results of research in different countries on the association between diabetes and CRC, I aimed to determine the strength of the association between diabetes and CRC in the NZ population with relation to diabetes duration and insulin use.

To address the aim, I focused on the following objectives:

1. Estimation of the CRC incidence rate ratios: in patients with diabetes relative to the non-diabetic population; in the entire NZ population; and in the population stratified by gender and by ethnicity (Q1).

2. Assessment of the effect of duration of diabetes on the CRC incidence rate ratios (Q2).

3. Calculation of the CRC incidence rate ratio in insulin users relative to the non-users with diabetes, and relative to individuals without diabetes (Q3).

4. Investigation of how the CRC incidence rate ratios in diabetic vs non-diabetic populations varies with age, stratified by gender or by ethnicity (Q4).

5. Estimation of incidence rate ratios for tumours located in the three anatomical sub-sites: in the proximal colon; in the distal colon; and in the rectum (Q5).

4.2 Methods

The methods section, in addition to the statements about study design and study population gives very detailed explanation of the variables used in the analysis. Because the preparation of the data set for statistical analysis included several steps, I also explain

the steps in details, including quality checks of the process. As the objectives were addressed by fitting several models, the subsection 4.2.3.5 ("Model fitting") is structured according to the study objectives with assigned names Q1–Q5 following the same order as in Section 4.1.5. The same order (from Q1–Q5) is used in presenting the results and also in the discussion section.

4.2.1 Study design and study population

This population-based cohort study covered the NZ population, 30-<90 years old, who were alive during the study period, i.e., between 1 January 2014 and 31 December 2018. A registration with diabetes mellitus on the Virtual Diabetes Register (VDR) between 1 January 2014 and 31 December 2018 was used as a marker of diabetes diagnosis. Patients with incident and prevalent diabetes were included in the study. Due to the lack of individual records for the non-diabetic population, to determine the size of the non-diabetic population, the corresponding count of the NZ population at 30 June each year, available from Statistics NZ, was used. The outcome was incident CRC diagnosis as registered in the NZCR (see Section 4.2.2.3).

4.2.2 Data

4.2.2.1 Data sources

Data sets used in this sub-study are listed in Table 4.1. As the data sets included many variables, but only a small portion of the variables were used in the analysis, the variables used are also provided in the table. The data sets were linked using the patients' encrypted NHI number.

Additionally, I used publicly available tabulations for the NZ population, obtained from

Source	Variables
VDR 2014–2018	Year, Gender, Ethnicity, Date of birth, Date of death, First outpatient date, First date of dispension, First lab result date, First inpatient date, NHI number
NMDS 1988–2018	Date, Event description, NHI number
NZCR 2014–2018	Gender, Ethnicity, Date of diagnosis, Age at diagnosis, Site of tumour, NHI number
Pharmaceutical Collection 2013–2016	Date of dispension, Chemical name, NHI number
Mortality Data 2018	Date of death, NHI number

Table 4.1: *Analysed data sets and variables in the data sets used for statistical analysis.*

Statistics NZ (Statistics NZ, 2020a). The tables included the count of NZ residents by 30 June each year from 2014–2018, stratified by sex, prioritised ethnicity and 1-year age brackets from 30 to 90 years.

The scales of certain variables used in the analysis set were different than in the original data sets. The differences are given in Table 4.2.

4.2.2.2 Exposure

The exposure was registration with diabetes. The exposed group consisted of patients with diabetes (included in VDR 2014–2018), who were 30 to 90 years old at the day of admission to the study and were free of CRC prior to diagnosis of diabetes. In this study, patients with prevalent and incident diabetes were included. Individuals were considered to have diabetes from the first registration date in VDR (first among: First outpatient date; First date of dispension; First lab result date; and First inpatient date). However,

Variable	**Source**	**Original resolution**	**Resolution in the analysis**
Age	Census	1 year	1 year
Age	NZCR, VDR	1 day	1 year
Calendar time	Census	1 year	1 year
Calendar time	NZCR, VDR	1 day	1 year
Gender	All	binary	binary
Ethnicity	All	categorical	binary (Māori/non-Māori)

Table 4.2: *Scales of variables in the original data sets and in the data sets used for statistical analysis.*

all individuals were considered to be at risk of CRC while having diabetes from 1 January 2014 or from the date of diabetes diagnosis, whichever came later (start of follow-up), to the date of CRC diagnosis, 90th birthday, death or to the end of the study (31 December 2018), whichever occurred first (end of follow-up). Diabetes was modelled as time-varying binary covariate: that is, patients with incident diabetes were considered non-diabetics before registration in VDR and, from the date of registration in VDR, as patients with diabetes.

The reasoning for the choice of diabetes mellitus rather than T2D as exposure is explained below.

Some studies investigating the association between diabetes and CRC calculated the increase in risk in patients with any type of DM (Carstensen et al., 2012; De Bruijn et al., 2014; Ohkuma et al., 2018), while others (Johnson et al., 2012; Harding et al., 2015) calculated the increased risk in T2D. When designing the study, I considered comparing the CRC risk in patients with T2D to the non-T2D population. However, the VDR data set does not distinguish between T1D and T2D. In order to classify patients as T1D or T2D it is necessary to use an algorithm based on a combination of variables which come

from different sources (NMDS and Pharmaceutical Collection). In NZ, such an algorithm was developed by McKergow et al. (2017) and used in NZ research, e.g. in Wheeler et al. (2019) to select patients with T1D. One of the data sources which the algorithm uses to select patients with T1D, is NMDS. Since many patients do not have records in NMDS (in VDR 2014–2017 32% of all patients did not have any NMDS record), and some have an inconsistent diabetes type recorded in NMDS, they cannot be classified reliably.

An additional problem with using NMDS data to select T2D is that nearly all CRC patients included in the VDR 2014–2018 had NMDS records, and therefore, the information included in NMDS data would usually suffice to establish diabetes type in patients with CRC. Thus, patients whose diabetes type could not be established would, almost exclusively, be non-CRC. If mostly non-CRC patients were excluded, it would lead to misclassification bias. Additionally, it would not be possible to test whether there is a difference between CRC risk in T1D and T2D in NZ patients using data available for this project. To analyse data from patients with T1D in NZ with satisfactory power, a very long follow up would also be required. Around 50 years of follow up would be required to reach one million person-years, which is the sample needed for such a study. An example of a study that assessed the risk of CRC in T1D and T2D is Harding et al. (2015), which analysed data from 950,000 patients from Australia with diabetes, including 81,000 patients diagnosed with T1D and a median follow up for T1D of 12 years. The study found no statistically significant difference in IRRs in T1D and T2D. The IRRs in the two groups were close to 1, and with a narrow 95% CI for T1, which suggests that in the Australian population the effect of T1D and T2D on incidence of CRC are similar. It is plausible that IRR in T1D and T2D in the NZ population will also be similar.

Since only approximately 5-10% of the diabetes population are patients with T1D and, among these, many may be misclassified, it makes little difference in the analysis, from a statistical point of view, if patients with T1D are included or not. In terms of implications in clinical practice, the following two issues can be considered:

1. Medical doctors will have different criteria for a diagnosis of T1D than any algorithm which I would use to select patients with T1D. In that case, it will be more practical to have a single model for the whole diabetes population; and

2. Concerning the use of the results for CRC screening criteria, it seems to be easier to stratify patients with diabetes using only one stratification factor: diabetes yes/no.

As reported by Atkinson et al. (2014), the correct differentiation of patients with T1D is nearly impossible. The criteria for diagnosis of T1D remain unclear and the diagnosis of T1D versus T2D, especially among adults, can be challenging. According to the authors, around 5-15% of adults who were diagnosed with T2D might actually have type 1 disease. If so, as many as 50% of actual patients with T1D could be misdiagnosed as T2D, meaning that the number of patients with type 1 disease is vastly underestimated. Also, a meta-analysis by Luo et al. (2016) reported that the risk of CRC in T2D was nearly identical to that of combined T2D and T1D. Taking such factors into consideration, for the purpose of this study, I decided to define the exposed group as patients diagnosed with any type of DM (except gestational diabetes).

4.2.2.3 Outcome

The primary outcome was diagnosis of **colorectal cancer**. Patients with incident CRC were identified from the NZCR years 2014-2018, using ICD-10 codes C18–C20. Cases with pre-existing CRC before 1 January 2014 (checked with NZCR registrations back to 1994) were excluded from the study.

Additionally, I defined three secondary outcomes based on anatomical location of tumours: diagnosis of **proximal, distal** and **rectal cancer**. Cancers located in the cecum (C18.0), appendix (C18.1), ascending colon (C18.2), hepatic flexure (C18.3), and transverse colon (C18.4) were categorised as proximal CRC; cancers located in splenic flexure

(C18.5), descending colon (C18.6) and sigmoid colon (C18.7) were categorised as distal CRC; and rectal included location in recto-sigmoid junction (C19) and in the rectum (C20). Patients with only overlapping (C18.8) and/or tumours with unspecified location (C18.9) were classified as unspecified location. Patients with synchronous tumours in different anatomical sub-sites were removed from the sub-site specific analysis.

4.2.2.4 Confounders and effect modifiers

Confounders: The following variables were considered as potential confounders: age (time varying covariate), gender, ethnicity, and calendar year (time varying covariate). I used prioritised ethnicity from two data sources, VDR and NZCR, to assign ethnicity for exposed participants. I defined ethnicity as follows: Māori when Māori was indicated in any record, non-Māori if any other ethnicity was indicated at least once, and otherwise unknown ethnicity. Because the data will eventually be compared to census data, which does not include unknown ethnicity, patients with unknown ethnicity were removed from the data set used in statistical analysis. Deprivation was also considered, but the covariate was not included in the statistical analysis because the appropriate tables, including count of population stratified by gender, ethnicity, age and deprivation, were not available from Statistics NZ and, therefore, tabulation by deprivation was not possible.

Effect modifiers: Age, gender, ethnicity and calendar year were considered as potential effect modifiers that affect both exposed and unexposed. Additionally, two effect modifiers that affect only exposed individuals, i.e., duration of diabetes and insulin use, were included in the analysis. Duration of diabetes (modelled as time-varying covariate) was tabulated in six intervals: 0–90 days, 91 days-<1 year, 1-<2 years, 2-<5 years, 5-<10 years and 10+ years. Insulin use was also modelled as time-varying covariate. Individuals were considered insulin users only if they had at least two redemptions of insulin within 6 months (following Carstensen et al. (2012)). In the statistical analysis they were

considered insulin users from the second date of insulin redemption in order to avoid immortal time bias. Patients with only one insulin redemption were analysed as non-insulin users.

4.2.2.5 Quality check of data

Quality checks of the following aspects of raw data were performed in order to check the quality of the data:

- consistency of gender, the date of birth and ethnicity between different data sources;
- plausibility of the order of the event times (e.g. no events before date of birth or after death);
- consistency between VDR and NMDS concerning the date of first and last discharge letters; and
- consistency of syntax of dates within each data set.

4.2.2.6 Data pre-processing

The original data sets include variables with information at the episode level, including cancer registrations (NZCR), yearly diabetes registration (VDR) and insulin dispension (PC). For each patient, there is more than one row in the original data sets, where each row relates to one episode. For the statistical analysis, for each of the data sets, a data set at the patient level was prepared. For patient-level data sets, information from all episodes was reduced to a single row for each patient (further referred to as patient-level data). In order to prepare the patient-level data set, the data sets listed in Table 4.1 were linked into a single patient-level dataset which was further extended with the following derived variables:

- **The date of diabetes registration** defined as the first record among: the first date of in/out-patient hospital visit, the first date of dispensing of medication to control diabetes, and the first date of laboratory result (from VDR). The date of registration of diabetes was used to calculate duration of diabetes for each exposed patient.

- **The date of insulin initiation** defined as the second date of insulin dispensing. To identify the dates of dispensing, each record was labelled as insulin or non-insulin, where insulin included any type of insulin plus accessories used to administer insulin.

- **The site of tumour location** for single tumours was defined as in section 4.2.2.3. Patients with multiple diagnoses, with a gap between diagnoses up to 90 days, were treated as having one CRC diagnosis with multiple tumours. The decision was based on the observation that in most patients with multiple tumours the gap between dates of diagnoses was less than 90 days (Figure 4.1). For patients with multiple CRC tumours, all anatomical locations of tumours were recorded in the patient-level data. The date of CRC diagnosis in patients with multiple tumours was defined as the date of the first diagnosis.

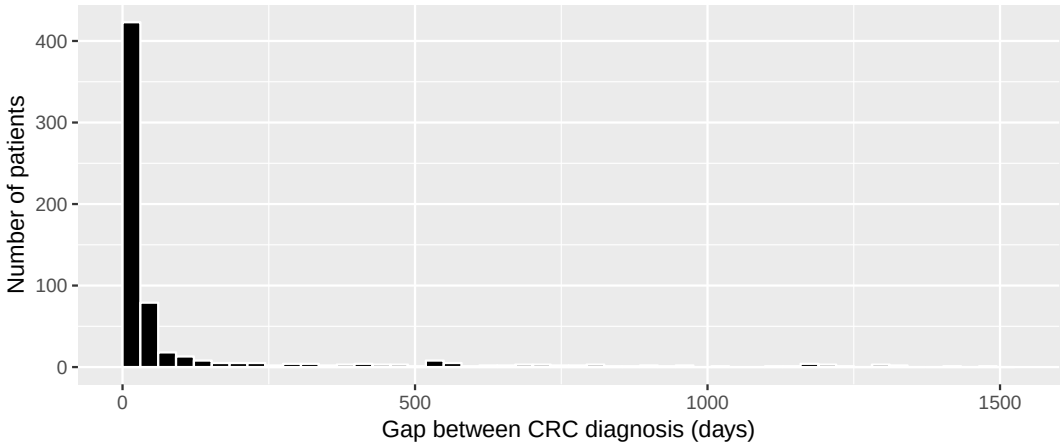


Figure 4.1: *Gap between diagnoses of multiple tumours.*

4.2.3 Statistical analysis

Cox regression and Poisson regression are the two models most commonly used in the earlier studies which investigated associations between diabetes and CRC. In this analysis, I used Poisson regression in order to incorporate population data available from Statistics NZ (structured in groups defined by one-year brackets stratified by gender and ethnicity), and also because Poisson regression handles well incidence data with multiple time scales (But et al., 2018), which were used in this analysis (age, calendar year, and duration of diabetes). The statistical analysis was based on a previously used framework for dealing with time-to-event data with multiple time-scales (Carstensen, 2007; Carstensen et al., 2012, 2014, 2016). The implementation was based on the *Epi* package for R (Carstensen et al., 2019) (for tabulation) and the *mgcv* package for R (Wood, 2017) (for model fitting). In this analysis, individual records were not available for the unexposed population and, therefore, data were tabulated prior to the model fitting. The details of the tabulation are explained in the next section.

4.2.3.1 Tabulations

The follow-up in patients with diabetes, measured in person-years (PY), was calculated for each individual following the principle illustrated in the Lexis diagram in Figure 4.2. General information about Lexis diagrams can be found in chapter 3 section 3.1.2.1. The Lexis diagram in Figure 4.2, for illustrative purposes, represents a snapshot from a full Lexis diagram for this sub-study. The snapshot includes only CRC cases restricted to the age groups 49–53 years. In the full Lexis diagram, each red line shows a patient's lifeline from the start of follow-up the end of follow-up, representing the patient contribution to the study (PY). The black dots represent dates of CRC diagnosis. The same principle was applied to calculate the number of PY for all patients with diabetes, with or without CRC. In the entire diabetes population, the total number of

PY and the number of CRC incidences were calculated for each combination of 1-year age bracket, gender, ethnicity, calendar year, duration of diabetes and insulin use (in the following referred to as *cells*). Duration of diabetes was tabulated in the following brackets: 0–3 months, 3months–1 year, 1 year–2 years, 2 years–5 years, 5 years–10 years, 10 years+. Insulin use was categorised as user/non-user (defined in section 4.2.2.4).

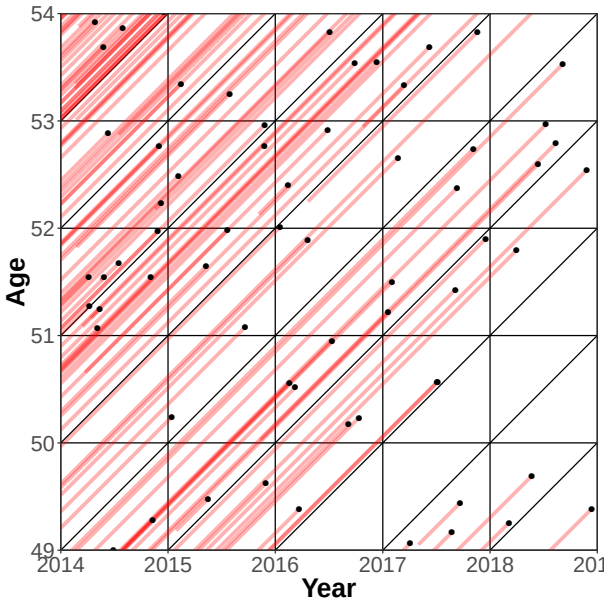


Figure 4.2: *Lexis diagram for a selected sample of CRC cases with diabetes.*

The contribution of each patient to the cell was calculated in time-dependent manners due to diabetes duration and insulin use being modelled as time varying covariates. This implies that, e.g., a patient who started to take insulin in the third year of follow up, would contribute as non-insulin user to each cell during the first two years of follow-up and then, from the time of starting the treatment, as an insulin user. The same principle applies to the duration of diabetes. That is, a patient would move from category to category as time since diabetes diagnosis advanced with the time of follow-up. In this way, before a diagnosis of diabetes a patient contributed person-years to the non-diabetes category. Just after the diagnosis of diabetes, the patient contributed to the 0–90 days since diagnosis category; as time of follow-up advanced, they might contribute to 91

days–1 year and possibly to the 1–2 and 2–5 years of diabetes duration categories.

In order to calculate the CRC incidence rate ratio (IRR) in diabetic vs non-diabetic populations, firstly the number of PY in each cell defined by: 1-year age bracket, gender, ethnicity and calendar year for the diabetic populations were calculated. In the non-diabetic populations, the number of PY in each cell was computed by subtracting the number of PY in the diabetic population from the size of the general population in that cell. The general population consisted of the count of the NZ population by 30 June each year, in 1-year age brackets, stratified by gender and ethnicity (Māori and non-Māori). Unlike in the diabetic population, the number of PY in the general population could not be calculated due to the lack of individual records and, therefore, the count of the population by 30 June of each year was used as a surrogate for PY. It could, in theory, cause a bias if, e.g., most deaths happened in spring. To investigate how serious the bias could be, the dates of deaths were plotted (the histogram is in Figure 4.3). As it was important for the analysis I present the histogram in this section. The distribution of deaths is nearly symmetrical around 30 June which gives confidence that, even if the bias is present, it should not have any substantial effect on the results.

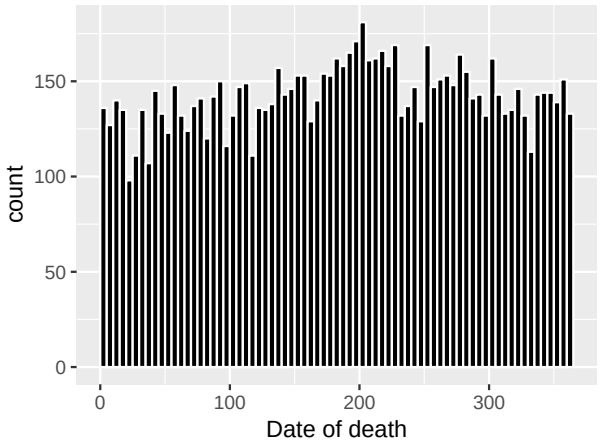


Figure 4.3: *Distribution of the day of the year of death for patients with diabetes.*

For the purpose of statistical analysis a table was constructed, where each row represents

a combination of 1-year age brackets, gender, ethnicity (Māori, non-Māori), calendar year and diabetes status (Table 4.3). Cells for diabetes population were also stratified by duration of diabetes and insulin use. Duration of diabetes was tabulated in six brackets as described above in section 4.2.3.1 One row in the table corresponds to a Lexis cell. For example, in the Lexis diagram (Figure 4.2) it can be seen that, in 50-year-old individuals with diabetes, in years 2014 and 2015, there was only 1 CRC patient that is indicated in the relevant column in Table 4.3.

Age	**Sex**	**Ethnicity**	**Year**	**Diabetes status**	**PY**	**CRC (n)**	**Duration**	**Insulin use**
50	F	Māori	2014	Diabetes	491.8	0	0-3 mths	No
50	F	non-Māori	2014	Diabetes	1511.6	0	0-3 mths	No
50	M	Māori	2014	Diabetes	500.9	0	0-3 mths	No
50	M	non-Māori	2014	Diabetes	1739.3	0	0-3 mths	No
50	F	Māori	2015	Diabetes	534.7	0	0-3 mths	No
50	F	non-Māori	2015	Diabetes	1564.4	**1**	0-3 mths	No
50	M	Māori	2015	Diabetes	540.1	0	0-3 mths	No
50	M	non-Māori	2015	Diabetes	1768.9	0	0-3 mths	No

Table 4.3: *Example of Lexis cells for patients 50 years old (partially in 2014 and 2015), with diabetes duration 0-90 days.*

4.2.3.2 Quality check of tabulation

I performed a quality check of the accuracy of the tabulation of the follow-up time and the total number of CRC cases included in 10 randomly selected Lexis cells for males with diabetes, as well as 10 cells for females with diabetes. The selection was made using the *sample* function in R.

The number of PY and CRC cases in each of the selected cells was compared to the calculated values for the same stratum from the data set at the patient-level. In case of any discrepancies, the pre-processing and tabulation was corrected until achieving a satisfactory agreement, i.e., a discrepancy in follow-up time of up to 1 day/patient was tolerated as it could result from, e.g., rounding off errors or leap years. However, with respect to the number of CRC cases, no discrepancies were tolerated, which additionally was checked by comparing the total number of CRC cases in the analysis set and in the raw data.

4.2.3.3 Exploratory analysis and sample description

To become familiar with patterns in the data, I carried out exploratory data analysis based on histograms for continuous variables, tables for categorical variables and scatter plots for relations between two continuous variables. The exploratory analysis was stratified by gender, ethnicity and diabetes status. Data for the exposed population and for patients with incident CRC were summarised using median and interquartile range (IQR) for continuous variables, and frequencies and percentages for categorical variables. Sample characteristics for patients with diabetes were stratified by gender and ethnicity. For CRC patients, characteristics were stratified by diabetes status. For the unexposed group, the number of PY and the number of CRC cases are reported. I restricted the analyses to patients with complete data on all variables included in the models. To assess whether the analysis was biased due to removal of patients with missing values, characteristics of patients with missing values were compared to those with complete data.

4.2.3.4 Statistical model

I estimated the CRC incidence rate ratios in diabetic populations compared to non-diabetic populations using Poisson regression models, with cells described in section

4.2.3.1, as observational units. The outcome was the number of CRC cases in each cell, which was assumed to be Poisson distributed with the expected number of CRC cases in cell i, λ_i, equal to the number of PY in cell i multiplied by the incidence rate for cell i (predicted using the covariates for cell i). To implement the model in R, the PY was included in the model as offset (log (PY)), which means that the coefficient of log (PY) was set to 1. The justification for the assumption of Poisson distribution is the small probability of CRC in an individual during the time described by a particular cell (see section 4.2.3.1). Further, I assumed that the joint effect of the exposure and other covariates is additive on the log scale. This means that the predicted incidence rate of CRC related to a combination of risk factors changes by a factor equal to the product of the effects of the individual risk factors (Greenland, 1989). The effect of age was modelled using thin plate splines as recommended by Wood (2003), where the number of degrees of freedom was chosen using generalised cross-validation as implemented in the *mgcv* package for R. Age was modelled on continuous scale using splines, as it resembles the natural relation between age and CRC incidence while avoiding overfitting the data (Wahba, 1990). Duration of diabetes was modelled as a categorical term (categories defined in Section 4.2.3.1). The fitted Poisson model had the formula in equation 4.1:

$$\begin{aligned} n_{CRC_i} \sim Poisson(\lambda_i)\ where:\ log(\lambda_i) = \beta_0 + spline(Age_i) + \beta_1 Sex_i + \\ \beta_2 Ethnicity_i + \beta_3 Diabetes\ status_i + \\ \beta_4 Calendar\ time_i + \beta_5 Insulin\ use_i + \\ \gamma_2 Duration_{i2} + ... + \gamma_6 Duration_{i6} + \\ log(PY_i) \end{aligned} \tag{4.1}$$

where: λ_i is the expected number of CRC cases in cell i

$spline(Age)$ is a cubic spline with 6 knots

$\beta_1 ... \beta_5$ are parameters to be estimated

$Duration_{i2-6}$ are indicator variables for duration intervals

and γ_{2-6} are corresponding parameters to be estimated

PY_i is the number of patient years in cell i

The model includes $log(PY)$ with coefficient fixed to 1, which follows from the equation 4.2:

$$IR = \frac{n_{CRC}}{PY} \implies log(n_{CRC}) = log(IR) + log(PY) \tag{4.2}$$

To compare regression coefficients between demographics and between tumour sites, I used the Z-test shown in equation 4.3 (Paternoster et al., 1998):

$$Z = \frac{\beta_1 - \beta_2}{\sqrt{(SE\beta_1)^2 + (SE\beta_2)^2}} \tag{4.3}$$

4.2.3.5 Model fitting

To choose confounders for the main model which were also included in all fitted models, I initially fitted a model with all potential confounders and all possible two-way interactions between age, gender, ethnicity and calendar year. Subsequently, the initial model was simplified using backward elimination based on Wilk's test. Only terms statistically significant at the 5% level were included in the final model. To answer the stated research questions, I fitted several Poisson regression models, using different predictors which were chosen beforehand based on *a-priori* hypotheses and depending on the question to answer. The fitted models, with reference to the specific research question, are specified below.

Q1: IRR in diabetes vs non-diabetes. After the selection of confounders, I fitted a simple model, assuming no effect of diabetes duration and insulin exposure, to obtain

a CRC incidence rate ratio between diabetic and non-diabetic populations which could be compared to the majority of previously published studies. Diabetes status entered the model as a time-dependent variable, allowing diabetes status to change during the follow-up if diabetes were subsequently diagnosed. Further, the model was extended with relevant interaction terms or covariates for answering specific research questions. To investigate the CRC incidence rate ratios separately in males, females, Māori and non-Māori, I fitted models with the interaction terms diabetes status:gender and diabetes status:ethnicity. The effect of calendar year on IRR was assessed by including an interaction term diabetes status:calendar year.

Q2: Effect of diabetes duration. In the next step, the CRC rate ratio in diabetes in relation to the non-diabetic population was modelled as a function of duration of diabetes. Duration of diabetes was modelled in a time-dependent manner, allowing patients to contribute to multiple categories. The effect of duration of diabetes was assessed additionally for both genders, and for Māori and non-Māori.

Q3: Effect of insulin use. To investigate whether the CRC incidence rate in insulin users differs from that in non-insulin users, the simple model was extended with the covariate *insulin use*. Insulin exposure was modelled as a time-varying covariate, allowing a change in patient status from non-insulin user to insulin user during the follow-up. Insulin users were individuals who redeemed at least two prescriptions, and, therefore, a patient before redemption of the second prescription, was analysed as a non-insulin user, and only from the time after the second prescription as an insulin-user. In the analysis including exposure to insulin, I followed patients from 1 January 2014 to 31 December 2016 due to availability of data from Pharmaceutical Collection.

Q4: Effect of age. To investigate whether there are any specific age groups within the NZ population with diabetes, stratified by gender or by ethnicity, with especially increased CRC incidence rates compared to those without diabetes, models with two-

way interactions (diabetes:age) and three-way interaction terms (diabetes:age:gender and diabetes:age:ethnicity) were fitted. The 95% CIs (for the IRR for each one year age group) were computed using 10,000 simulations, using the *gratia* package for R (Simpson, 2018a), as explained in Simpson (2018b).

Q5: IRRs for anatomical sub-sites. The investigation of the effect of diabetes on CRC incidence with respect to the anatomical location of tumours was assessed by fitting three models with the outcomes: number of distal CRCs; number of proximal CRCs; or number of rectal CRCs in each Lexis cell defined by age, gender, ethnicity, calendar year and diabetes status. Patients with multiple tumours in different anatomic locations were removed from this analysis.

4.2.3.6 Model validation

To validate the fit of Model 1 (Table 4.7) to the data, I used two methods. Firstly, the number of CRC cases predicted by the model were plotted agains the observed numbers. The predicted and observed values were calculated for 20 quantile brackets defined by the linear predictor from the model. The values were compared using a Chi-squared test (Wood, 2002). The second check was carried out using the *gam.check* function from *mgcv* R-package to obtain a quantile-quantile plot for the deviance residuals (Augustin et al., 2012) and the fit was assessed visually.

Calendar year was modelled as having a linear relationship with the logarithm of the CRC incidence rate. The assumption was tested by fitting two versions of Model 1 (Table 4.7), with calendar year coded as categorical variable and coded as a continuous term. The Likelihood Ratio Test was used to test the hypothesis that there were no differences in the predictive accuracy of those two models.

4.2.3.7 Sensitivity analysis

The following sensitivity analyses were carried out to assess the sensitivity of the inference to various factors:

1. To assess how much a possible detection bias could affect the overall CRC rate ratio, I carried out two analyses by restricting the analysis based on Model 1 (Table 4.7) to individuals who were diagnosed with CRC at least 90 days, and at least 180 days after diabetes diagnosis.

2. To assess whether the study was affected by factors which could bias the results towards the null association, I investigated whether inclusion of prevalent diabetes in the analysis could cause the estimated IRR to be lower than in many studies. This was done by carrying out an analysis limited to incident diabetes only, that is to patients who had a first record related to diabetes from 1 January 2014 to 31 December 2018. Firstly, I estimated the overall IRR for the whole sample of incident diabetes. Inclusion of only incident diabetes would most likely introduce a detection bias; therefore, to assess the size of the detection bias, I estimated the IRR for those diagnosed with CRC later than 90 days from diabetes registration, and finally for those diagnosed with CRC 180 days after diabetes registration. Additionally, by fitting Model 2a using incident diabetes, I visually assessed the pattern of IRR in relation to the duration of diabetes.

3. To assess the possible impact of an incorrect date of diabetes diagnosis on the overall IRR, patients with a VDR registration date within 30 days after CRC diagnosis were treated as diagnosed with diabetes before CRC diagnosis. The crude IRR was calculated and compared to the crude IRR from the original analysis.

4. To investigate whether the analysis was sensitive to the adjustment for confounders, the crude IRRs, the IRRs adjusted for age and gender, and the fully adjusted IRRs

(for age, gender, ethnicity and calendar year) were computed.

All statistical analysis was carried out using R version 3.6.2. A significance level of 5% was used throughout.

4.2.3.8 Determination of age for CRC screening in diabetes

Although not within the scope of this study, I propose a simple way in which the results of the analysis performed in this study could be used for the stratification of the NZ population for CRC screening. Rather than using age as the only criterion for eligibility (as in the current policy of 60–74 years as a universal threshold for all), the age threshold for population-based screening could depend on diabetes status. The age for patients with diabetes would be determined in the following steps:

1. Determination of the IR in the general population at age 60 years could be done by fitting a Poisson regression model with age as the only predictor.

2. Determination of the age in the diabetic and non-diabetic populations that corresponds to the IR found in step 1 could be done by fitting a Poisson regression model with age, diabetes status and interaction age:diabetes.

The decision whether to include the interaction age:diabetes in the above model was based on the results from Model 4a in Table 4.7.

4.3 Results

4.3.1 Quality of data and accuracy of tabulation

The overall data quality was good. A few discrepancies in gender and date of birth were resolved by assuming the information in NZCR to be correct. In VDR, there were inconsistencies in the dates of events for some patients across five years (inpatient event, outpatient event, laboratory event and pharma event) which were used to determine the date of diabetes diagnosis. The date of diabetes registration was assigned to the first date of any event across all five years VDR. I found a surprising pattern in the overall number of events related to diabetes (inpatient event, outpatient event, laboratory event and pharmaceutical event). As can be seen in Figure 4.4, there was a lower number of events in years 2009–2012 than in adjacent years, with a sharp increase in 2013. This is unlikely to reflect the real number of diabetes events, but it is rather an artefact related to the administration. As a consequence, some patients who had their first record in 2013 could have the duration of diabetes misclassified (see 4.2.2.6). The misclassification would most likely affect duration of diabetes longer than 1 year. A patient who, at a given time-point, had a calculated duration of, e.g., 1.5 years (bracket 1–2 years) could, in reality, belong to the next bracket 2–5 years, or to the following 5–10 years. The effect of this possible misclassification was checked by carrying out a sensitivity analysis using only incident diabetes (presented in section 4.3.7 point 2).

In the performed quality check for the accuracy of the tabulation, I achieved satisfactory agreement between raw data sets and the data set used in the analysis, which confirmed that the R-code for preparation of the data set was reliable. There were very small discrepancies in numbers of PY, with no patients with a discrepancy of more than 1 day (during the whole follow-up).

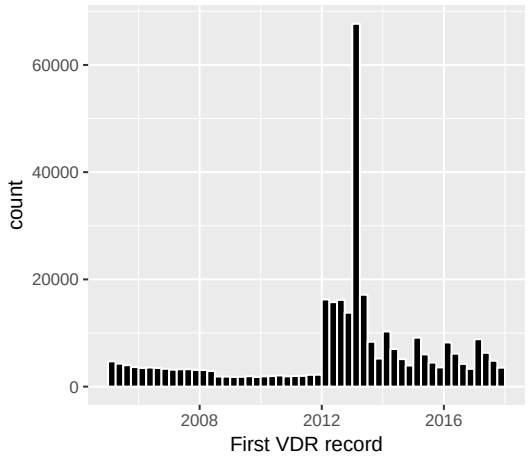


Figure 4.4: *Distribution of first registrations in VDR.*

4.3.2 Participants

Potentially eligible participants were the entire NZ population between 1 January 2014 and 31 December 2018, who were 30–90 years old during the study, equal to 13.7 million PY. Among those, during the study period, 14,606 patients were diagnosed with incident CRC, but only 14,437 were analysed, as patients with unknown ethnicity (n=169, 1.2%) were removed from the analysis.

Between 1 January 2014 and 31 December 2018, 341,103 individuals who were alive at 1 January 2014 were registered in VDR with diabetes mellitus. Of those, 313,322 satisfied the inclusion criteria. However, 2612 (0.8%) patients with diabetes were removed from the analysis due to unknown ethnicity. The detailed process of the selection of exposed individuals for statistical analysis is shown in Figure 4.5.

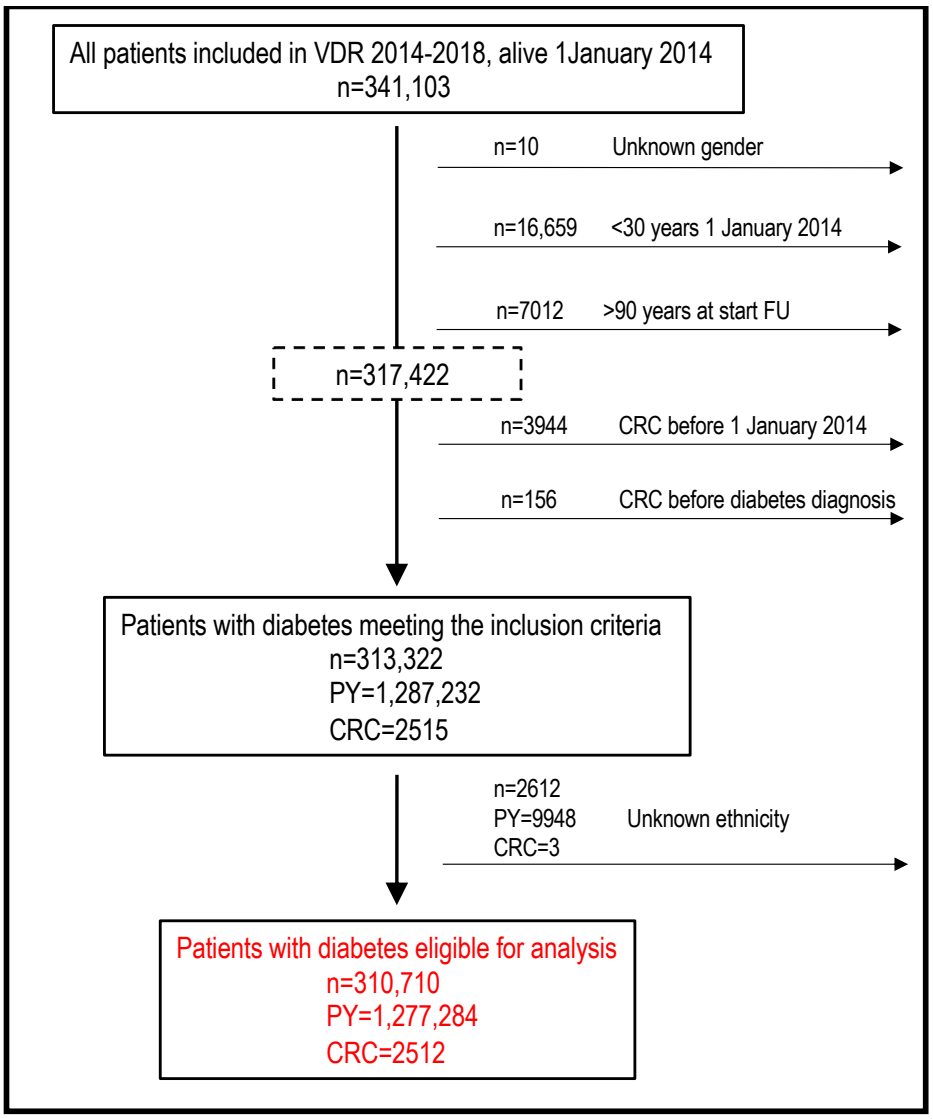


Figure 4.5: *Selection of the study population with diabetes eligible for analysis.*

Characteristics of patients with diabetes with known ethnicity, and those who were removed due to unknown ethnicity, are in Table 4.4. Compared to patients with known ethnicity, patients with unknown ethnicity were more likely to be males, less likely to have CRC, and lived in less deprived areas. Females with unknown ethnicity were older than females with known ethnicity [median age 71 years (IQR; 57, 80) and 62 years (IQR; 52, 71) respectively].

	Unknown ethnicity			Known ethnicity		
Variable	**All**	**Males**	**Females**	**All**	**Males**	**Females**
n	2612	1886	726	310,710	162,542	148,168
% Males	–	(72,2 %)	–	–	(52.3 %)	–
Age						
median (IQR)	64 (54-74)	62 (53-70)	71 (57-80)	62 (52-71)	62 (52-71)	62 (51-73)
Deprivation						
%>5	57.3	56.7	58.8	62.9	61.4	64.5
CRC						
%	0.11	0.05	0.28	0.80	0.90	0.70

Table 4.4: *Comparison of demographics of the diabetes population with missing ethnicity and without missing ethnicity.*

4.3.3 Descriptive statistics

In total, data from 310,710 patients with diabetes, corresponding to 1,277,284 person-years, with 2512 patients registered with incident CRC during the study, were analysed. Of those, 59.9% had full follow-up (5 years). Among patients with diabetes, 52.3% were males. The median age at the start of follow-up did not differ between males and females. Both genders were equally likely to use insulin during the follow-up (based on data from 2014 to 2016). 17.7% of the diabetic population were Māori. In comparison, of the whole analysed population, 10.7% were Māori. Median age at start of follow-up in Māori was 57 years, compared to 63 years in non-Māori. Māori were slightly less likely than non-Māori to be insulin users (19.7% vs 21.7%). Characteristics of patients with diabetes are in Table 4.5. The unexposed group (individuals without diabetes) consisted of 12,422,463 PY with 11,925 incident CRC patients diagnosed during the study period (2014–2018).

		Gender		Ethnicity	
Variable	**All**	**Male**	**Female**	**Māori**	**Non-Māori**
	n=310,710	n=162,542	n=148,168	n=55,111	n=255,599
CRC during follow-up					
n (%)	2512 (0.8)	1468 (0.9)	1044 (0.7)	269 (0.5)	2243 (0.9)
Age at start of follow-up					
median (IQR)	62 (52-71)	62 (52-71)	62 (51-72)	57 (48-66)	63 (53-72)
Deprivation					
(% >5)	62.9	61.4	64.5	83.2	58.5
Insulin use (%)	20.1	19.9	20.2	21.7	19.7

Table 4.5: *Characteristics of the study population with diabetes.*

Among the 14,437 analysed patients with CRC, 17.4% had diabetes. In comparison to non-diabetics, CRC patients with diabetes were slightly older (median age at CRC diagnosis 74 vs 71 years), and were more likely to be male, Māori, and to live in more deprived areas. Characteristics of patients with incident CRC, separated for individuals with and without diabetes, are in Table 4.6.

		Exposure to diabetes	
Variable	**All**	**Yes** (17.4 %)	**No** (82.6 %)
Gender			
Female	6791	1046	5745
Male	7646	1466	6180
% Male	53	58.4	51.8
Age			
Median	71	74	71
(IQR)	(62, 79)	(66, 80)	(60, 79)
Ethnicity			
Māori	1015	269	746
non-Māori	13,422	2243	11,179
% Māori	7.0	10.7	6.3
Deprivation			
Low (<6)	6856	1024	5832
High (>5)	7554	1486	6068
% High	52.4	59.2	51
Tumour characteristics			
Tumour site (%)			
Proximal colon	37.5	38.1	37.3
Distal colon	28.3	28.1	28.4
Rectal	34.1	33.6	34.2
TNM stage of CRC (%)			
I	20.2	20.9	20.1
II	15.5	16.0	15.4
III	21.8	20.5	22.0
IV	20.1	19.7	20.1
Unstaged	22.5	22.9	22.4

Table 4.6: *Characteristics of incident CRC cases stratified by diabetes status.*

4.3.4 Model specifications

The following confounders were included in all fitted models: age; gender; calendar year; ethnicity; and the interaction ethnicity:calendar year. The effect of calendar year on CRC incidence was different in both ethnicities ($p<0.01$ for the interaction) and, therefore, the interaction was included in all models. Table 4.7 gives the specification of fitted models with reference to the research question, which was addressed by fitting a specific model.

Research Q	Model	Outcome	Predictors in the model
Q1	Model 1	CRC	**Age**, **Sex**, **Diabetes status**, **Ethnicity**, **Calendar year**, **Ethnicity:Calendar year**, **offset(log(PY))**
	Model 1a	CRC	Model 1 + **Diabetes status:Sex**
	Model 1b	CRC	Model 1 + **Diabetes status:Ethnicity**
	Model 1c	CRC	Model 1 + **Diabetes status:Calendar year**
Q2	Model 2a	CRC	Model 1 + **Duration of diabetes**
	Model 2b	CRC	Model 1 + **Duration of diabetes:Sex**
	Model 2c	CRC	Model 1 + **Duration of diabetes:Ethnicity**
Q3	Model 3	CRC	Model 1 + **Insulin use**
Q4	Model 4a	CRC	Model 1 + **Diabetes status:Age**
Q4	Model 4b	CRC	Model 1 + **Diabetes status:Age:Sex**
Q4	Model 4c	CRC	Model 1 + **Diabetes status:Age:Ethnicity**
Q5	Model 5a	Proximal	Model 1
	Model 5b	Distal	Model 1
	Model 5c	Rectal	Model 1

Table 4.7: *Specification of the fitted models with relation to the addressed research question.*

4.3.5 Results from fitted models

The results of statistical modelling are presented below with the reference to each of the research objectives (Section 4.1.5).

Q1: IRR in diabetes vs non-diabetes. The overall CRC incident rate in the NZ population with diabetes was increased by 13% with IRR=1.13 (95%CI; 1.08, 1.18) compared to the population without diabetes. In males, the rate ratio was a bit higher than in females (IRR=1.16 vs IRR=1.08 respectively), but the difference between genders was not statistically significant (p=0.16 for interaction diabetes:gender). In Māori, there was no association between diabetes and CRC [IRR= 0.95 (95% CI; 0.83, 1.09)]. The rate ratio in non-Māori [IRR= 1.15 (95% CI; 1.09, 1.20)] was statistically significantly increased compared to Māori individuals (p=0.01 for the interaction term diabetes:ethnicity). Calendar year was found to be a statistically significant effect modifier of the association between diabetes and CRC, with IRRs decreasing by a factor of 0.96 (95% CI; 0.94, 0.99) per calendar year.

Q2: Effect of diabetes duration. The effect of the duration of diabetes is shown in Figure 4.6. The rate ratios (for the whole population with diabetes) were highly increased shortly after diagnosis of diabetes (from 0 to 90 days) with IRR=2.55 then decreased during the first year to IRR=1.15, reaching an IRR of 1.09 by around 5 years duration of diabetes, and approaching the incidence rate in the non-diabetic population around 10 years after diabetes diagnosis (Table 4.8). The pattern is very similar in males, females, Māori and non-Māori. There is an indication that detection bias (shown by higher rate ratio for duration up to 90 days) might be higher in males and in Māori but, as can be seen, the 95% CIs are very wide, which means we are unable to draw any informed conclusions about differences between strata.

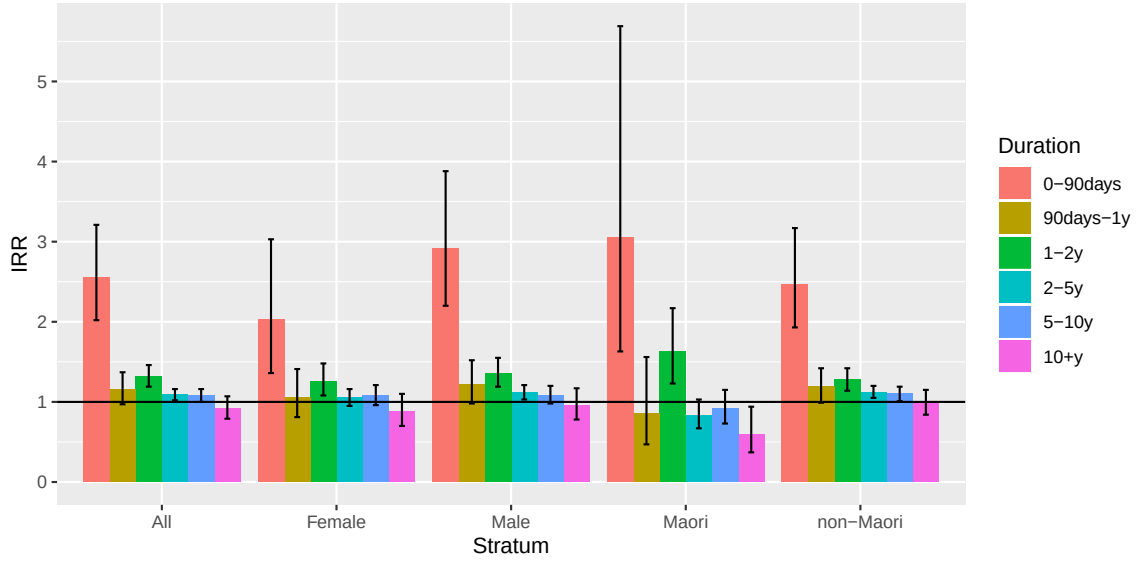


Figure 4.6: *The effect of duration of diabetes on CRC incidence rates ratios for all, males, females and Māori with 95% CIs shown by the error bars. A null effect is indicated by the black horizontal line.*

Q3: Effect of insulin use. The flow chart in Figure 4.7 shows the count of the population used in the analysis of the association between insulin use and CRC (data from only three years of VDR). In total, data from 283,858 patients with diabetes, corresponding to 735,156 person-years, were analysed. During the three-year period, 8719 patients were diagnosed with incident CRC, of which 1518 (17.4%) had diabetes, and 315 (20.8%) of those were insulin users.

The overall CRC incidence rate ratio based on data from three years was nearly the same as based on five years' data, IRR=1.14 (95%CI; 1.08, 1.21). In insulin users, CRC incidence was increased by 32% when compared to the non-diabetic population, and by 20% when compared to individuals with diabetes who do not use insulin. In non-users of insulin, compared to the non-diabetic population, the rate was also increased slightly, by 10%. All values are presented in Table 4.8.

Q4: Effect of age. The increase in CRC incidence in diabetes was not the same

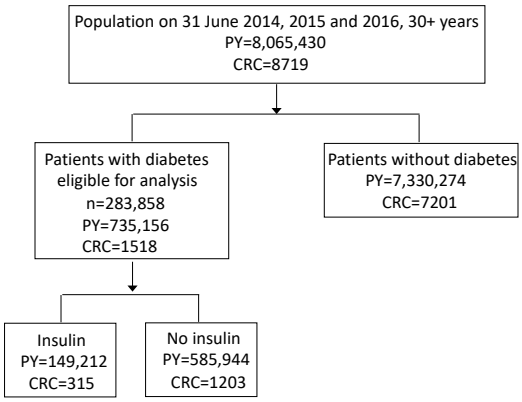


Figure 4.7: *Sample used for investigation of the effect of insulin use on CRC incidence.*

across the whole age range. Rate ratios in diabetic vs non-diabetic populations were increased only in those younger than 75 years. However, as can be seen in Figure 4.8, in individuals 30 to just over 40 years old, the 95% CI is wide due to the small number of CRC patients with diabetes, and includes the null value. In patients older than 75 years, CRC incidence rates in those with diabetes were similar to the rates in those without diabetes.

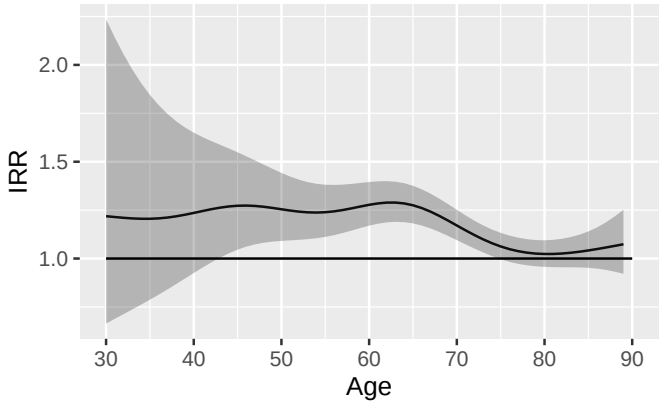


Figure 4.8: *Association between diabetes and CRC incidence as a function of age, from the model adjusted for gender, ethnicity and calendar year for all.*

The analysis including three way interactions revealed that the increase in CRC incidence in patients with diabetes applied only to non-Māori (Figure 4.9), and was statistically significantly increased only in those aged 45–75 years (p=0.045 for the three way interaction age:diabetes:ethnicity). There is an indication of an increased incidence rate of CRC in young Māori (<50 years) with diabetes, compared to those without diabetes, but, due to the small number of Māori, the 95% CI is wide and includes the null value. With respect to gender, the effect of age on the association between diabetes and CRC did not differ statistically significantly between males and females (p=0.44 for the three way interaction age:diabetes:gender).

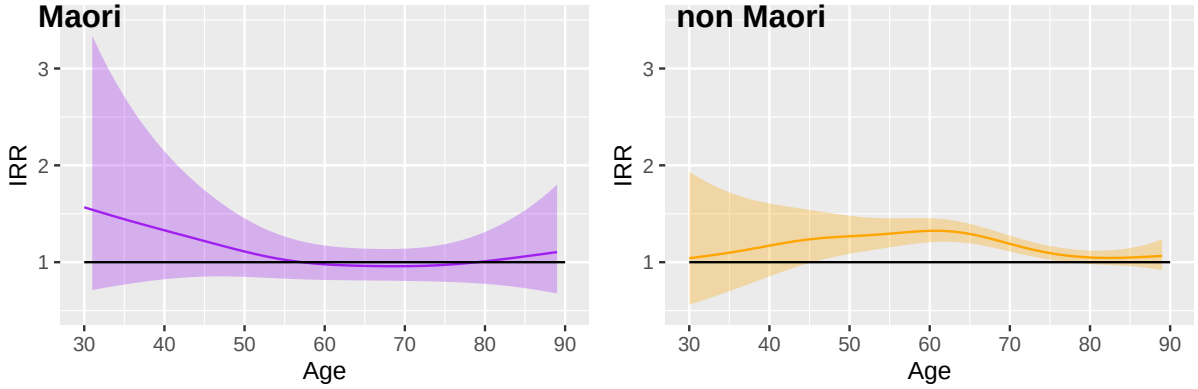


Figure 4.9: *Association between diabetes and CRC incidence, in Māori and non-Māori, as a function of age, from models adjusted for gender and calendar year.*

Q5: IRRs for anatomical sub-sites. In the analysis with sub-site specific CRC incidence as outcome, the incidence rate ratios of proximal, distal and rectal tumours in patients with diabetes vs non-diabetics were similar, with the highest IRR=1.17 for rectal cancer. There was no statistically significant difference between the IRRs for distal and proximal cancers (p=0.87, Z-test), distal and rectal cancers (p=0.36, Z-test), and proximal and rectal cancers (p=0.25, Z-test). In the sub-site analysis by gender, I found no statistically significant interaction between diabetes and gender for any of the three sub-sites (p values are in Table 4.8, which provides all IRRs with 95% CIs for all

fitted models).

Question	Model†	Effect of diabetes	IRR (95% CI)
Q1	1	overall	1.13 (1.08, 1.18)
Q1	1a	by gender	**Males** 1.16 (1.09, 1.22)
			Females 1.08 (1.02, 1.16)
Q1	1b	by ethnicity	**Māori** 0.95 (0.83, 1.09)
			non-Māori 1.15 (1.09, 1.20)
Q1	1c	by calendar year‡	**2014;** 1.21 (1.12, 1.30)
			2015; 1.17 (1.11, 1.23)
			2016; 1.13 (1.08, 1.18)
			2017; 1.09 (1.03, 1.15)
			2018; 1.05 (0.97, 1.13)
Q2	2b	by duration	**0-0.25 y:** 2.55 (2.02, 3.21)
		as categorical	**0.25-1 y:** 1.15 (0.97, 1.37)
			1-2 y: 1.32 (1.19, 1.46)
			2-5 y: 1.09 (1.02, 1.16)
			5-10 y: 1.08 (1.00, 1.16)
			10+ y: 0.92 (0.79, 1.07)
Q3	3	by insulin use	**Insulin vs non-diab** 1.32 (1.18, 1.48)
	-		**Insulin vs non-insulin** 1.20 (1.06, 1.36)
			Non-insulin vs non-diab 1.10 (1.04, 1.17)
Q4	4a		Figure 4.8
	4b		p=0.44 for the interaction
	4c		Figure 4.9, p=0.03 for the interaction
Q5	5a	by site	**Proximal** p=0.14*: All 1.10 (1.02, 1.18)
			Males 1.17 (1.05, 1.30)
			Females 1.05 (0.95, 1.15)
	5b		**Distal** p=0.43*: All 1.11 (1.02, 1.20)
			Males 1.14 (1.02, 1.27)
			Females 1.06 (0.92, 1.22)
	5c		**Rectum** p=0.08*: All 1.17 (1.08, 1.26)
			Males 1.22 (1.12, 1.34)
			Females 1.06 (0.93, 1.21)

*P value for the interaction diabetes:gender

†Models specification in Table 4.7

‡The linear effect based on modelling assumption

Table 4.8: *The incident rate ratios for diabetes vs non-diabetes from fitted models.*

4.3.6 Model validation

There was no evidence for the disagreement between the fitted (model 1) and observed incident rates. The chi-square test did not reject the hypothesis (p=0.48) that the number of CRC cases in each of the 20 brackets, defined by quantiles of the linear predictor, were Poisson distributed, with λ equal to the exponential of the linear predictor, multiplied by the number of person-years for each bracket. The model fit is illustrated in Figure 4.10.

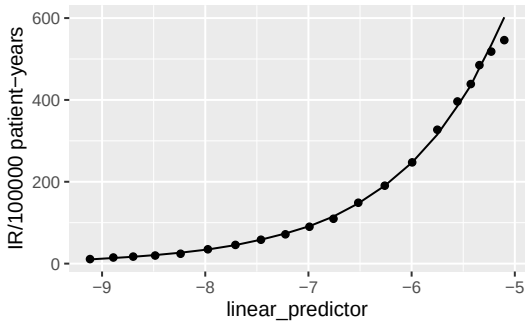


Figure 4.10: *Model predicted IR (curve) compared to empirical IR (dots) based on Model 1.*

The actual distribution of the deviance residuals across Lexis cells was very close to the theoretical distribution, and the scatter plot shows good model fit to the data (Figure 4.11).

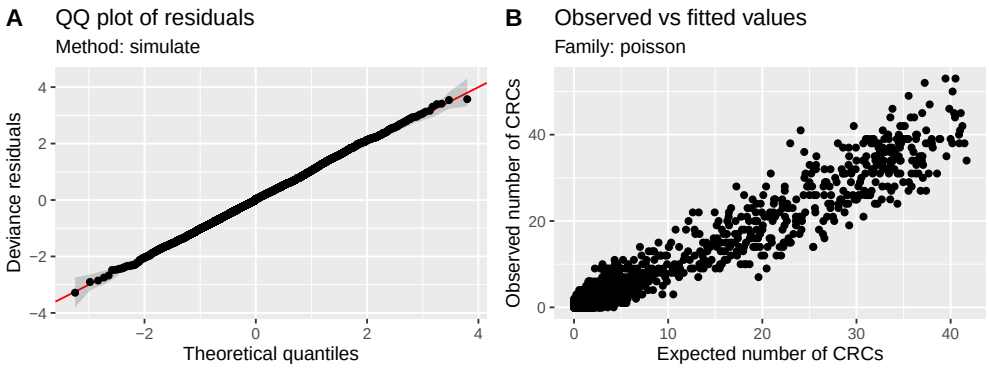


Figure 4.11: *Model fit across Lexis cells.*

The assumption of the linear effect of calendar year on CRC incidence rate was not rejected (p=0.34) (Figure 4.12).

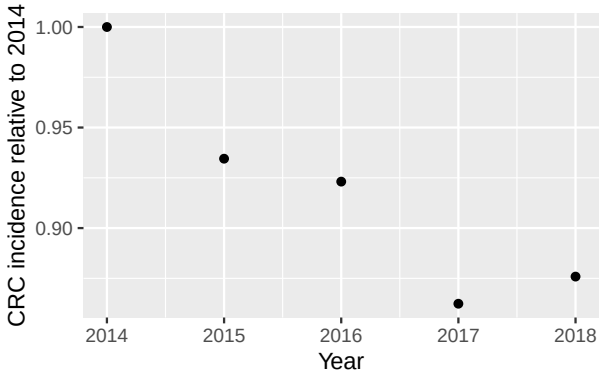


Figure 4.12: *Effect of calendar year on CRC incidence, measured as IR relative to IR for year 2014.*

4.3.7 Sensitivity analysis

1. In the sensitivity analysis for the assessment of detection bias, carried out using only data from patients with CRC diagnosis at least 90 days after diabetes registration, the estimated rate ratio for the duration 0.25–10+ years was IRR=1.11 (95% CI; 1.06, 1.16) and, for the duration 1 year–10+ years was 1.10 (95% CI; 1.05, 1.15). The discrepancy between those estimates and the overall IRR=1.13, which was calculated for the whole study period (0–10+ years), shows that only a very small detection bias was present in this study.

2. The IRR estimated in the analysis restricted to incident diabetes [IRR=1.27 (95% CI; 1.16, 1.40)] was higher than the IRR estimated in the analysis based on incident and prevalent diabetes. This analysis introduced a large detection bias. The IRR for patients with incident diabetes diagnosed with CRC between 90 days and 5 years after diabetes diagnosis was much lower [IRR=1.16 (95% CI; 1.04, 1.29)]. Further,

in the analysis of incident diabetes with CRC diagnosed between 180 days and 5 years from diabetes registration, IRR was 1.12 (95% CI; 1.00, 1.25), a similar effect size as in the main analysis, which included prevalent and incident diabetes.

3. A possible incorrect ordering of the events is not likely to have any substantial influence on the IRR, as only five patients had diagnosis of diabetes within 30 days after diagnosis of CRC. Analysis including those five patients as diabetics did not change the crude IRR (IRR changed from 2.048 to 2.053).

4. The crude IRRs were higher than the IRRs from the model adjusted for age and gender (Table 4.9). The additional adjustment for ethnicity, calendar year and interaction ethnicity:calendar year had very little impact on the values of IRR.

Stratum	CRC diab	CRC non-diab	PY diab	PY non-diab	IR diab	IR non-diab	crude IRR	adjusted IRR	fully adjusted IRR
All	2512	11,925	1,277,416	12,422,463	196.6	96.0	2.05 (1.96, 2.14)	1.12 (1.07, 1.17)	1.13 (1.08, 1.18)
Male	1468	6180	664,336	5,897,593	221.0	104.8	2.11 (1.99, 2.23)	1.15 (1.09, 1.22)	1.16 (1.09, 1.22)
Female	1044	5745	613,080	6,524,870	170.3	88.0	1.93 (1.81, 2.07)	1.08 (1.01, 1.16)	1.08 (1.02, 1.16)
Māori	269	746	227,047	1,263,343	118.5	59.0	2.01 (1.74, 2.31)	0.94 (0.82, 1.08)	0.95 (0.83, 1.09)
Non-Māori	2243	11,179	1,050,369	11,197,555	213.5	99.8	2.14 (2.05, 2.24)	1.14 (1.09, 1.20)	1.15 (1.09, 1.20)
2014	526	2440	231,597	2,410,322	227.1	101.2	2.24 (2.04, 2.47)	1.20 (1.11, 1.29)	1.21 (1.12, 1.30)
2015	488	2363	245,959	2,440,060	198.4	96.8	2.05 (1.85, 2.26)	1.16 (1.10, 1.22)	1.17 (1.11, 1.23)
2016	504	2398	257,985	2,479,505	195.4	96.7	2.02 (1.83, 2.22)	1.12 (1.07, 1.17)	1.13 (1.08, 1.18)
2017	506	2288	269,513	2,520,576	187.7	90.8	2.07 (1.87, 2.28)	1.09 (1.03, 1.15)	1.09 (1.03, 1.15)
2018	488	2436	272,360	2,572,000	179.2	94.7	1.89 (1.71, 2.09)	1.05 (0.98, 1.14)	1.05 (0.97, 1.13)

Table 4.9: *Crude and adjusted incidence rates of CRC in the NZ population (2014-2018) exposed and unexposed to diabetes, and model-based IRRs by demographics and calendar year.*

4.3.8 Estimated age for CRC screening in diabetes

The incidence rate of CRC for 60-year-old individuals (the lower bound of age for screening) from the general population equals to 94.7 CRC cases per 100,000 patient-years. The same IR for patients with diabetes corresponds to an age just below 57.5 years, which would be proposed as a screening age for those with diabetes (Figure 4.13 A)

when screening according to the risk. According to the model, in patients who take insulin, an age of 55 years corresponds to CRC incidence equal to the incidence in the general population at age 60 years (Figure 4.13 B).

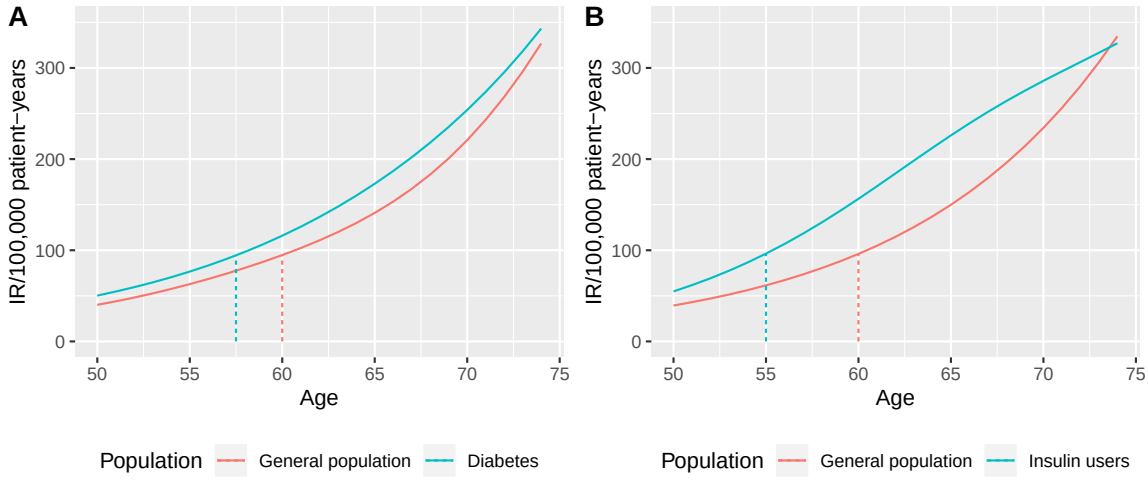


Figure 4.13: *Determination of screening age for patients with diabetes (a) and for patients who use insulin (b). The green curves are based on the model with only age as a predictor, the red curves are based on models including interactions age:diabetes status.*

4.4 Discussion

The primary focus of this sub-study was to investigate the association between diabetes and CRC in the NZ population, both overall, and in strata defined by gender, and by ethnicity. The study was not designed to investigate a causal effect of diabetes and/or insulin use on CRC developement. Instead, it aimed to provide reliable estimates of the strength of the association between diabetes and CRC in the NZ population - estimates that are relevant for clinical practice as a diagnostic criterion when investigating for the presence/absence of CRC. In the next sections, I will summarise the main findings and compare these findings to the literature. Subsequently, I will discuss the validity of the

results and explain their relevance to clinical practice. I will then discuss strengths of the research and problems related to the study design before, finally, proposing some ideas for future research.

4.4.1 Summary and discussion of main findings

The estimated CRC incidence rate ratio in diabetic vs non-diabetic populations was slightly increased [IRR=1.13 (95%CI; 1.08, 1.18)]. The estimated effect size was smaller than reported in meta-analyses (Larsson et al., 2005; Jiang et al., 2011; Starup-Linde et al., 2013; Wu et al., 2013; Luo et al., 2016; Sacerdote and Ricceri, 2018) and smaller than recently reported in similarly developed countries such as Australia (Harding et al., 2015), Canada (Sikdar et al., 2013), or the UK (Peeters et al., 2015). The weaker association in the NZ population could be partially due to the high rate of undiagnosed diabetes in NZ compared to other countries (Coppell et al., 2013). Also, the methodological differences between studies with respect to the lack of investigation of detection bias, the use of too-crude age brackets (e.g. 5 years), different inclusion criteria, or modelling time-varying covariates as fixed terms, could all possibly play a role.

Interestingly, ethnicity was found to be an effect modifier of the association between diabetes and CRC, with the IRR in non-Māori (IRR=1.15) increased by approximately 20% compared to IRR in Māori (IRR=0.95). The difference between the effect of diabetes on CRC incidence in Māori and non-Māori could perhaps be explained in the light of the findings from a recent Swedish study (Ahlqvist et al., 2018), which identified five different types of diabetes: it might be that diabetes in Māori has different genetic characteristics than diabetes in non-Māori, and that the association with CRC may be different in different types of diabetes. This is, however, a speculative explanation, as there is no published data on genetic characteristics of diabetes in Māori and non-Māori.

Also surprising and difficult to explain was the finding of a lack of association between

diabetes and CRC among Māori [IRR= 0.95 (95% CI; 0.83, 1.09)]. A similar phenomenon - i.e., differences between ethnic groups with respect to the diabetes-CRC association - have also been reported in a US study by He et al. (2010). The US study found no association between diabetes and CRC in Native Hawaiians (RR=0.89, 95% CI; 0.62, 1.27), while the association was found in other ethnicities. He et al. did not propose any explanation as, in the first place, the wide 95% CI makes it difficult to assess if there is true absence of the association in Native Hawaiians. In contrast, in my study the 95% CI is narrow, and it is very unlikely that the association in Māori could be positive.

If the true IRR in Māori is close to one, the results would imply that diabetes is not associated with CRC in Māori. This finding might be explained by important confounders which the analysis did not include such as lifestyle or choice of medication for controlling diabetes.

Alternatively, the rate ratio in Māori could be biased by the possibly higher number of underreported Māori individuals with diabetes compared to non-Māori with diabetes. According to the Auckland Diabetes Heart and Healthy Survey, the ratio of undiagnosed diabetes in patients from Auckland was higher among Māori (Sundborn et al., 2007). It is not very likely, however, that many of the Māori individuals with undiagnosed diabetes were CRC patients, as during the process of CRC diagnosis, diabetes most likely would be caught and registered. I therefore would not expect a different rate of underreported Māori and non-Māori with diabetes among CRC patients. The undercounting of Māori with diabetes in my analysis would rather cause overestimation of the IRR,[3] and therefore a higher underreporting of diabetes would not cause the lack of the association between diabetes and CRC in Māori.

[3]Conceptually, IRR in Māori is a ratio of two IRs: IR in Māori with diabetes where the denominator is "PY for Māori with diabetes", which would be too low due to undiagnosed diabetes causing overestimated IR in Māori with diabetes; and IR in Māori with denominator "PY for Māori without diabetes", which would be too high causing underestimated IR in Māori without diabetes. CRC cases with/without diabetes in numerators we consider correct. Therefore the IRR would be overestimated.

The inclusion of duration of diabetes in the analysis enabled the identification of an interesting pattern: namely, that the IRRs were highly increased in the first 3 months after diabetes diagnosis, decreased strongly within the following year, and remained slightly increased with duration of diabetes longer than 2 years. There was an indication that, in males and Māori, the IRRs were higher (vs females and non-Māori respectively) in the first 90 days of diabetes duration, which could indicate that males and Māori underuse medical services. However, this is rather speculative, as 95% CIs were very wide due to the small number of CRC patients with diabetes in duration brackets 0–90 days (especially in Māori). The described pattern, similar to what was reported in the Danish population (Carstensen et al., 2012) and in Canadians from British Columbia (Johnson et al., 2011), indicates the presence of a detection bias which is a result of the increased medical surveillance in patients newly diagnosed with diabetes, leading to the diagnosis of already present, but not yet diagnosed, CRC. In my study, the observed detection bias had only a small influence on the value of IRRs, which could be seen from the results of sensitivity analysis that discarded data from patients with CRC diagnosed within 90 days from diabetes diagnosis (IRR=1.11).

Insulin use has also been found to be an effect modifier of the association between diabetes and CRC. The incidence rate of CRC in insulin users was increased by 20% relative to non-users of insulin with diabetes. The increase in CRC incidence in insulin users vs non-users has been reported in earlier studies (Yin et al., 2014) and, as suggested by others, could be due to insulin exposure being a proxy for general frailty, which increases cancer risk in general, rather than the effect of insulin itself (Yang et al., 2004; Limburg et al., 2005).

Earlier studies did not commonly investigate whether age is an effect modifier in the association between diabetes and CRC. In papers which I reviewed, I have not found such an investigation, and therefore I cannot relate my results to what others found. The result of the analysis which investigated whether the diabetes-CRC association is the

same across the whole age span (30–90 years), showed that, within the NZ population, the association holds only in non-Māori younger than 75 years. This result can help partially to explain the disparities with some studies, as some studies included participants only at certain age, like e.g. He et al. (2010) who included only 45–75 years old participants. If the analysis carried out in my study had included only those younger than 75 years, the estimated IRR would be higher than the estimated IRR when individuals up to 90 years of age were included (this can be seen in Figure 4.8).

Finally, the analysis by anatomical sub-site showed that, in the NZ population, diabetes was associated with a similar increase in CRC incidence in all three analysed sub-sites. The IRR for rectal cancer had the highest value; however the difference was not statistically significant from the IRRs for proximal or distal cancers.

4.4.2 Comparison with earlier studies

Below I compare my results to studies published from 2010 onwards. This comparison is motivated by my finding in this sub-study, that the association between diabetes and CRC became weaker with time. It is possible that similar pattern is present in other populations, and therefore comparison of the results with results of older studies would be less informative, as it would not be clear if the differences are partially due to the effect of time.

Comparison of my findings to what others found, clearly shows that the association between diabetes and CRC in the NZ population is somehow weaker than that reported in similarly developed countries. A question however arises: is there a real difference in CRC incidence rates in diabetes within different populations, and therefore the effect size found in this study is truly lower than in most Western populations, or is this difference simply an artefact of different designs and different ways of reporting? On the one hand, as observed in this sub-study, there are differences between the diabetes-CRC association

by ethnicity within one country, which supports the hypothesis of differences between countries (González et al., 2017). On the other hand, some earlier studies have been affected by methodological problems related to statistical analysis, particularly by not including the duration of diabetes in the analysis (Johnson et al., 2012; Carstensen et al., 2012). Carstensen et al. have argued that "... analyses ignoring diabetes duration are meaningless, as they average over an influential variable (duration) with weights that are strongly dependent on the study design". Comparison is therefore not straightforward, especially since studies use very different methodologies and forms of statistical analysis, often not explained in sufficient detail to enable other researchers to judge the differences between important details in the analysis.

To explore this issue, firstly, I will compare my results to some studies which showed higher estimates possibly due to problems with study designs. The association in my study is much lower than, for example, in a study based in Italy (Valent, 2015) with respect to both colon and rectal cancers. The Italian study reported age- and gender-adjusted HRs in its analysis, with diabetes treated as time-varying covariate, for colon cancer to be 1.48 (95% CI; 1.36, 1.62), and for rectal cancer to be 1.26 (95% CI; 1.10, 1.45). However, their finding may have been influenced by an inappropriate adjustment for age. The participants were followed for 12 years, but the analysis was adjusted for the age at the start of observation. According to the paper's definition of the start of observation, a patient who got diabetes, for example, five years later was analysed as being 5 years younger compared to the actual age at diabetes diagnosis. As the incidence (or hazard) of CRC due to age increases during a 5-year period, the estimated HR was due not only to the diabetes effect; a part of the estimate was due to the effect of older age on CRC risk. The magnitude of the increase in CRC incidence due to age can be seen in Figure 4.13 A. For example, other things being equal, in 65-year-olds the incidence was increased by almost 50% compared to 60-year-old individuals.

The association found in this sub-study is also lower than from a case-control study

by Peeters et al. (2015), who analysed data from the British Clinical Practice Research Datalink with median follow-up of 4.5 years and reported HR=1.26 (95% CI 1.18, 133). The analysis included several predictors; however, duration of diabetes was not included. The authors acknowledged that detection bias might have affected their results, leading to an overestimation in the primary analysis, but it is not known to what extent. Although the sensitivity analysis carried out in my study showed that detection bias only moderately increased the overall IRR, detection bias could affect different studies to a larger extent. This can be seen from estimates by Johnson et al. (2011), discussed in the next paragraph, as well as from a more recent analysis of data from the Dutch population by de Jong et al. (2017), who reported a detection bias of a high magnitude for colon cancer [HR=1.4 (95% CI; 1.10, 1.70) before correction for detection bias; HR=1.2 (95% CI; 0.96, 1.60) after correction]. It is therefore possible that the true value of the increase in CRC incidence in the diabetic population vs the non-diabetic population studied by Peeters et al. (2015), could actually be similar to that in the NZ population.

As the above two examples show, it is possible that some of the differences between results of my study and results of other studies could be explained by differences in methodology and biases, and that unclear reporting could also play a role (Vandenbroucke et al., 2014).

On the other hand, the reported higher estimates from different populations can be genuine. A recent study by de Kort et al. (2017), who analysed data from newly diagnosed diabetes patients from the southern region of the Netherlands, and who assessed the effect of detection bias, similarly reported especially increased HR in the first 6 months from diabetes diagnosis. In their analysis, which excluded the first 6 months of follow-up, CRC risk remained increased in patients with diabetes [HR=1.3 (95% CI 1.2, 1.5)] compared to the estimate which included the first 6 months [HR=1.4 (95% CI 1.2, 1.6)]. Although de Kort et al. (2017) found higher detection bias than my study, the estimated HR in patients with diabetes in the Dutch study was higher than the IRR in the NZ

population.

Sikdar et al. (2013) provides another example of a study which reported a much higher increase in CRC incidence in the diabetic population (in males HR=1.38, in females HR=1.52) than what I found. This cohort study, which analysed data from Canadian patients living in Newfoundland and Labrador, included a long follow up (over 10 years) using incident diabetes only. Patients who were diagnosed with CRC within 1 year from diabetes diagnosis were removed from the analysis, thereby addressing a possible detection bias. There are no particular methodological differences between Sikdar et al. (2013) and my study that could explain the different estimates. The most likely explanation is a real difference between these two populations - Canadians from Newfoundland and Labrador, and New Zealanders - which could be due to differences in lifestyles and environmental factors.

Comparison with studies which have investigated the temporal relationship between diabetes and cancer risk shows that the pattern of the CRC incidence related to duration of diabetes found in this study was very similar to that reported in Johnson et al. (2011), Carstensen et al. (2012), and Harding et al. (2015). In an analysis of data from British Columbia, Johnson et al. (2011) estimated the HR for patients with diabetes duration 0–3 months as 2.89 (95% CI; 2.22, 3.75), for 3 months–10 years as 1.15 (95% CI; 1.05, 1.25) and for 0 months–10 years as 1.24 (95% CI; 1.14, 1.35). This result shows the effect of detection bias (the difference between rate ratios of 1.24 and 1.15). As we can see, the estimated HR for 3 months-10 years is only slightly higher than my estimate of the IRR in the NZ population. The Australian study by Harding et al. (2015) reported a similar pattern; however, the detection bias in the first 3 months after diabetes diagnosis was weaker (SIR=1.47) than in my study (IRR=2.55).

Ballotari et al. (2017) is an example of a study that included duration of diabetes in the analysis, but not modelled as a time-varying covariate. That is, their patients were

assigned to one fixed duration bracket at the start of the study, although in reality many patients contributed to more than one duration bracket. Possibly due to the different analysis, Ballotari et al. (2017) reported a pattern opposite to that found in my study. The IRR was lowest in the 0–2 year bracket, increased for up to 10 years duration of diabetes, and then lowered again for a duration longer than 10 years. Due to differences in methodology, it is difficult to compare my results to Ballotari et al. (2017).

The results from my study were also similar to the results from a recent prospective study in China which included over 0.5 million of patients with diabetes (Pang et al., 2018). Although Pang et al. reported a slightly higher estimate [RR=1.18 (95% CI; 1.04, 1.33)], they reported a similar association for rectum and colon similar to what I found in this study. The results from this sub-study also did not differ substantially from the results of two Australian studies: the above-mentioned study by Harding et al. (2015), who reported the overall SIR for T2D to be 1.18 (95% CI; 1.15, 1.20) and for T1D SIR=1.21 (95% CI; 1.06, 1.37); and the recent study by Kelty et al. (2019) who analysed data from individuals aged 18–69 years and reported diabetes to be associated with an increased risk of CRC in a multivariable model that controlled for several factors such as family history of CRC, alcohol and tobacco use and comorbidities, with OR=1.17 (95% CI; 1.02, 1.34).

With respect to the effect of diabetes on the incidence of proximal, distal and rectal cancers, the three effects which did not differ statistically significantly in this sub-study, my results are different to what was reported by some other studies. Two Dutch studies - Overbeek et al. (2019) and de Kort et al. (2016) - reported males with T2D to have higher risk of distal tumours than males without diabetes, while Dutch females with T2D, especially at older ages, were reported to have an increased risk of proximal tumours (Overbeek et al., 2019).

4.4.3 Validity of the study results

The main biases which could be a threat to the internal validity of the study (detection bias and immortal-time bias) were addressed in the study design and statistical analysis. Detection bias was mitigated by the inclusion of prevalent diabetes and was assessed by the inclusion of duration of diabetes in the analysis. Immortal-time bias was addressed by counting person-years for insulin users only from the date of the second redemption of insulin. However, the internal validity of the results could be distorted by a possible non-representativeness of VDR, which could cause misclassification bias. The representativeness of VDR depends on eligibility criteria for inclusion, which was beyond my control. I was not able to conduct any sensitivity analysis to investigate the magnitude of the impact which the potential misclassification bias had on the overall IRR, due to the lack of data for those who had diabetes, but were not recorded in VDR. Therefore, the results are valid only under the assumption that the patients included in VDR are a representative sample of the NZ population with diabetes.

As the fitted model was only controlled for age, gender, ethnicity and calendar year, it may seem at first glance that the internal validity of the results was impaired by unmeasured confounding by factors such as obesity and sedentary lifestyle; especially given that this information can be elicited by physicians during the appointment for investigation of presence/absence of CRC. Firstly, such data are not available at population level in NZ, and secondly, even if the data were available, it might not be appropriate to include such confounders in the model. This is because CRC develops over a very long period (5, 10 or even 20+ years) and therefore a patient's recent lifestyle and obesity status may be misleading. A patient who has currently appropriate weight, could be exposed to obesity a long time before and the past exposure could affect the initiation and development of CRC. It is therefore, arguably, more accurate to use diabetes status as a proxy for various unmeasured long-term exposures, as was done in this sub-study.

With respect to the external validity, the results of the study are not applicable to other countries due to the differences between IRRs in different ethnic groups in NZ, which makes it likely that IRRs will also differ between countries.

4.4.4 Strengths

The main strength of this sub-study was the population-based nature, which assured that the results were not affected by selection bias. Further, the analysis of administrative data sets made it possible to analyse data from a large number of individuals (ensuring narrow CIs) within a short period (a three-year PhD project) at low cost.

Another strength was the choice of Poisson model for data analysis with Lexis cells as observational units, which is an appropriate choice when it comes to the analysis of time-to-event data with several joint time scales (age, calendar year, duration of diabetes). Additionally, the model enabled modelling time-varying covariates (diabetes status and insulin use) in time-dependent manners, which captures the real life situation - that is: a patient can contribute firstly to the follow-up as a non-diabetic person and, after diabetes diagnosis, as a diabetic individual. The same applies to the use of insulin. It can be seen that this analysis does not require making a choice at which time-point a patient has to be included as having diabetes or as an insulin user. In studies which do not model time-varying covariates as such, it could be the start of the follow-up or the date of CRC diagnosis; however, neither choice mirrors the realistic situation, which might change with time.

An additional strength was modelling age on a continuous scale using splines - not commonly used in studies on the diabetes-CRC association. It is important to model age appropriately, as age is a major confounder in the association between diabetes and CRC. Further, in clinical practice decisions are often taken with relation to exact age, e.g. screening policy. Therefore, if data are available with fine time resolution, assessing

CRC incidence as a function of the exact age in the case of CRC is more useful.

Modelling insulin use in time-depending manners allowed the analyses to include patients who started to take insulin during the follow-up: excluding those patients would have led to an under-estimated IRR by introducing misclassification bias. This is because use of insulin is related to more advanced forms of diabetes, and thus more ill patients, who are more likely to have CRC, would be analysed as non-users.

As the follow-up in this study was not very long (maximum 5 years follow-up), the inclusion of prevalent diabetes in addition to incident diabetes was a strength, rather than a weakness. Inclusion of patients with prevalent diabetes avoided the estimated CRC incidence rate ratio in diabetes being affected substantially by detection bias. As discussed by Carstensen et al. (2012), inclusion of prevalent diabetes can induce possible problems with, e.g., the assessment of the duration of diabetes, if the full diabetes history for patients with prevalent diabetes is not known. However, following the argument: with respect to duration of diabetes, I had full information going back to 1988 (only few patients had a longer history, but they belonged to the longest duration i.e., 10+ years, where the analysis does not distinguish between 10, 12 or 14 etc. years).

Inclusion of the duration of diabetes in the analysis enabled the assessment of the effect of detection bias. Detection bias would lead to overestimated IRR, which could cause unnecessary anxiety in patients with diabetes, as well as increasing the burden on the health system if policymakers decided to include diabetes status in the pathway for CRC diagnosis.

4.4.5 Limitations

The study has several limitations that have to be acknowledged.

Firstly, diabetes status was assigned from VDR, a virtual register, which is less accurate

than a register based on the actual diagnosis of diabetes. The sensitivity of VDR reported by the MoH, defined as the proportion of patients diagnosed with diabetes who are registered in VDR, is 87% (Ministry of Health NZ, 2018c). This figure implies that some patients with diabetes were included in our analysis as patients without diabetes, which could cause a slight underestimate of IRR.

Secondly, since MoH expects a 40% higher prevalence of diabetes than the actual VDR registrations, a high number of patients with undiagnosed diabetes (∼100,000, including those who do not know that they have diabetes and those who were diagnosed but are not included in the VDR) were analysed as not exposed to diabetes. If there were no systematic pattern in the undiagnosed population, and VDR were a random sample of patients with diabetes in NZ, it would have very little influence on the results. However, if there is a selection bias within VDR, i.e. if a substantial proportion of unregistered individuals with diabetes belongs to a group with a particularly low or high CRC risk, the true CRC incidence rate ratios, overall and in strata, would differ from those calculated in my study, and the effect could go in either direction. An example could be a population with undiagnosed diabetes characterised by poorly managed hyperglycaemia, which, according to hyperglycaemia hypothesis, is positively associated with CRC (Chang and Ulrich, 2003).

Third, patients who were removed due to missing ethnicity were not representative of the population with diabetes; in particular, they were much less likely to have CRC (Table 4.4). It could in principle cause the estimated IRR to be overestimated; however, taking under consideration the very low number of removed patients (0.8%), the value of the overestimation will be very small. I have chosen to perform a complete case analysis because probabilistic imputation of missing ethnicity based on other information available would be very complicated, as the Lexis procedure require each patient to be assigned 100% to one cell, while the procedure of probabilistic imputation would assign patients partially to cells (Enders, 2010). Multiple imputation was possible, but the

tabulation is already computationally very demanding, and multiple imputation would make it much worse. Taking under consideration the small size of the bias, I decided to perform complete case analysis.

Fourth, the large sample size enabled stratification of the analysis by gender, ethnicity and multiple duration brackets, which increased the possibility of a statistically significant finding by chance only, as I used 5% significance level to assess the significance of the interactions. I did not correct for multiple testing in accordance with the recommendation by Greenland (1989). In the author's view, the correction for multiple testing in epidemiological studies would prevent researchers from finding new ideas which, after further research, might lead to better understanding of many health related topics.

A further potential limitation in my study is the lack of controlling for important potential confounders such as BMI or physical activity level, which is a result of the sub-study design and the use of administrative data which do not contain that type of information. The results of this sub-study are also affected by possible biases related to the data quality, which are explained in the next two paragraphs.

Despite the overall good data quality, the lack of systematically collected information about a clear date of diabetes diagnosis was the most important issue that could lead to biased estimates of IRR. Due to possible misclassification of dates of diabetes registration, some patients with a diagnosis of diabetes and CRC within a short time window, could be recorded in VDR shortly after CRC diagnosis when, in fact, they have been diagnosed with diabetes before CRC diagnosis. They were thus analysed as non-diabetics while they were, in fact, diabetes patients. Based on the sensitivity analysis, in which patients with a diabetes diagnosis up to 30 days after CRC diagnosis were treated as CRC patients with diabetes, I found that the misclassification of dates did not impact the results as the crude IRR was nearly identical to the crude IRR from the main analysis (point 4 Section 4.3.7).

The lower number of registrations in VDR around years 2008–2011, which was followed by a highly increased number in 2013, could potentially affect the results. Most likely, more patients were diagnosed with diabetes in years 2008–2011 than the number of registrations in VDR shows, but due to, e.g., changes in the medical or administrative procedures, the registrations were included in VDR only in years later than the actual diagnosis of diabetes happened. Additionally, in 2012 new guidelines for the diagnosis of diabetes in NZ were released, which presumably led to the peak in the numbers of diabetes registrations in 2013 (Ministry of Health NZ, 2012). As a result, some patients could have a longer diabetes duration than the duration which was assigned to those patients. Specifically, the group with duration longer than 1 year would be expected to be affected and, therefore, the estimates of IRR which include those with 1+years of diabetes duration could be slightly affected. However, this inaccuracy did not influence the CRC incidence rate ratios from models that did not include duration of diabetes.

4.5 Conclusions

In this study I found that NZ patients with diabetes have moderately increased incidence rates of CRC compared to the non-diabetic population (IRR=1.13), however only non-Māori younger than 75 years. The overall IRR=1.13 was only minimally affected by the detection bias.

Duration of diabetes influenced the estimated incidence rate ratios. The IRR was substantially increased in the first three months after diabetes diagnosis which can be explained by increased medical surveillance of patients shortly after diabetes diagnosis.

Carrying out statistical analysis based on data from the NZ population was important as the IRR turned out to be lower than in many other countries. Modelling age using splines allow to identify an interaction between age and diabetes; an important finding that may

partially explain differences between studies due to the different inclusion criteria with respect to age. In addition, the analysis revealed different strength of the association between diabetes and CRC in Māori and non-Māori.

The substantially increased CRC incidence in insulin users (IRR=1.32 vs non-diabetes) is likely a proxy for the overall health and extent of diabetes. It is not showing a causal effect of insulin treatment on CRC incidence rates.

This section finishes sub-study 2 which explored if CRC incidence rates in the NZ population with diabetes differ from the rates in individuals without diabetes. Similarly like sub-study 1, this sub-study was a population based study and modelled incidence rates in the whole NZ population. The following chapter (Chapter 5) presents an analysis of data collected from a single hospital in order to fit a model for CRC risk in individual patients.

Chapter 5

Model for CRC risk in patients referred to secondary care

This chapter describes a sub-study which developed a statistical model for CRC risk in individual patients referred to secondary care for further investigation for the presence/absence of CRC. For those patients, specialists have to decide whether a referral will be accepted or rejected. A model for calculation of CRC risk in those patients could be a helpful tool in such decision-making[1].

5.1 Introduction

Early diagnosis of CRC in symptomatic patients can be a challenging process and, as explained in Chapter 2, Section 2.4.2, can involve several steps. In one of the final steps of the diagnostic pathway in NZ (Figure2.2), gastroenterologists or general surgeons have to decide based on a referral made by a GP whether a patient should be investigated further for the presence/absence of CRC. The investigation usually would be carried out by performing a colonoscopy, which poses two main issues: societally, the issue of possible inefficient use of colonoscopy resources (Stamm et al., 2020) in NZ; and, individually, the small risk of serious health complications (Juillerat et al., 2009). For both reasons, it would be beneficial to perform colonoscopies on patients who have a sufficiently high CRC risk to justify a colonoscopy (Adelstein et al., 2010, 2011).

The choice of patients for further investigation using colonoscopy, however, is not trivial. In the first place, there is no clear agreement among researchers about how predictive of CRC different symptoms are (Ford et al., 2008; Jellema et al., 2010). In order to establish a consensus on which symptoms would improve the yield of positive colonoscopies, the European Panel on the Appropriateness of Gastrointestinal Endoscopy (EPAGE) developed recommendations for the selection of patients for colonoscopy based on symptoms, age, family history of CRC, and history of earlier colonoscopy (Vader et al., 1999) [updated in 2008 (EPAGE II) Juillerat et al. (2009)]. However, the EPAGE experts also

[1]Preliminary results from this sub-study were presented in Hirsz et al. (2019)

sometimes disagreed on the importance of specific symptoms for the selection of patients for colonoscopy (Juillerat et al., 2009).

Further, the referral letter from the GP may include symptoms that are not specified as high-risk in guidelines, or may list a combination of low-risk symptoms, and e.g., information about family history of CRC. The complex process of combining relevant information about a patient together with presented symptoms is a process that doctors must often carry out in their medical practice to improve the accuracy of the assessment of the CRC risk (Ford et al., 2008). Because statistical models or algorithms offer strategies for analysing multifactorial relationships, this process of combining several pieces of information in clinical practice could be assisted by using diagnostic models (Ford et al., 2008; Adelstein et al., 2010, 2011; Williams et al., 2016).

The existing literature has indeed suggested that statistical models and algorithms which can combine several factors, such as patients' symptoms, demographic data, comorbidities and test results, improve the accuracy of predicting CRC risk for individual patients, compared to symptoms alone, which have a poor predictive values for CRC (Ford et al., 2008; Adelstein et al., 2011; Williams et al., 2016). There is also some evidence to suggest that some of the statistical models reviewed by Williams et al. (2016) show better discrimination than the existing referral guidelines (for a model based on secondary care data, see Selvachandran et al. (2002)).

Indeed, such concerns are understandable, considering that earlier models (Bellentani et al., 1990; Selvachandran et al., 2002) were criticised for their too low specificity, which results in performing too many colonoscopies on patients with low CRC risk (Ford et al., 2008). However, the weighted numerical score proposed by Selvachandran et al. (2002) could e.g., identify 91% of patients who actually had CRC (sensitivity) with corresponding specificity of 62%. In other words, using the score, only 38% of actually given negative endoscopic investigations would be performed, while 91% of CRCs could still be found. In

comparison, using NHS guidelines, both the sensitivity and the specificity would be lower (86% and 52% respectively). Actually, the specificity of 62% which shows the potential savings in the proportions of performed, but negative, colonoscopies, does not have to be so high, as any specificity higher than zero represents a savings compared to the current practice.

Following my argument, let us assume the goal is to improve the current practice, e.g., to reduce the number of unnecessary colonoscopies. In current practice, per definition, the sensitivity is 100% and the specificity is 0%, if, as in Selvachandran et al. (2002), endoscopy was used as the reference test, and therefore all patients included in the analysis actually had received an endoscopic investigation. Thus, predictive models have to provide high sensitivity, as it is not acceptable to use a model which will lead to missing more than a very small proportion of patients with CRC who were diagnosed in current practice. If this is coupled with the prevention of unnecessary colonoscopies, even at the level of 10%, the saved unnecessary colonoscopies would be a gain (even a specificity of 10% would be fully acceptable).

Ford et al.'s (2008) critique of low specificity of such models is, therefore, not necessarily adequate in all situations: the required specificity and sensitivity depend on the intended goal for a specific test or investigation (e.g. better management of resources) as well as on the study design (Pepe et al., 2003; Akobeng, 2007; Turner et al., 2019). As can be seen on the receiver operating characteristic (ROC) curves presented by Selvachandran et al. (2002) if the specificity were lowered to 46% percent, the sensitivity (which is the important measure in the case of models that could help to select patients with high risk without missing too many diagnoses of already existing and diagnosable CRCs) would go up (in this case, to around 99%, as shown in Table 4 in Selvachandran et al. (2002)).

A later study by Adelstein et al. (2010) presented a model fitted using data from Australian patients referred for colonoscopy. The model gave very promising results. The

authors calculated that, using their model, and performing only 60% of actually performed colonoscopies, 95% of patients from their cohort still would be diagnosed, an improvement of the current practice comparable to the score offered by Selvachandran et al. (2002).

All of the above models administered questionnaires to elicit information from patients about the variables included in their models. Although such questionnaires give detailed information about patients' symptoms, this is not the information which is available to the specialists when they assess a referral for acceptance or rejection. Therefore, an alternative way of eliciting symptoms would be extraction of symptoms from the clinical notes included in GPs referrals, as in Hsiang et al. (2013). Clinical notes can include coded and uncoded information. The uncoded information included in free-text fields is a valuable specification of a patient's disease, but the extraction of information from a free-text field written by GPs can be problematic and challenging as the information included in free-text comments is often ungrammatical, with unclear terms and abbreviations (Patrick and Asgari, 2010; Koeling et al., 2011).

Some authors elicited symptoms from doctors' notes by manual extraction (Hsiang et al., 2013). However, analysis of large data sets within a short period of time, as well as updating the analysis with new data, would be more efficient when using an automated extraction of symptoms, instead of a manual procedure (Meystre et al., 2008). There is an available algorithm for the extraction of symptoms related to ovarian cancer designed by Koeling et al. (2011) which also covers four symptoms related to CRC. The authors provide recall (sensitivity) for combined coded and free-text symptoms extraction, and additionally for the extraction of coded information alone. From those two numbers the calculated recall of the extraction from free text is as follows: 69% for bloating, 73% for abdominal pain, 74% for diarrhoea and 96% for constipation. A systematic review of algorithms for automated extraction of information from electronic health records by Ford et al. (2016) presented a median recall across all studies of 31% (also calculated

from provided recalls for coded and combined extraction). The much lower recall is likely due to the use of electronic health records that are of lower quality than the British GP Research Database used by Koeling et al. (2011).

To summarise, predictive models for the calculation of CRC risk have been shown to have a potential for assisting specialists in the decision-making process. If the model based on NZ data could achieve a reduction in performed colonoscopies similar to Adelstein's model (40%) and, at the same time, diagnose nearly all patients (95%), it seems to be an attractive opportunity to help with the management of the scarce resources for colonoscopies in NZ. To the best of my knowledge, the associations between CRC and symptoms included in free-text notes in e-referrals to the hospitals have not previously been studied in NZ; hence research in this area is needed. Hsiang et al. (2013) evaluated the so-called "Auckland scoring system" using symptoms present in CRC patients in free-text notes; however the study did not include data from patients without CRC. In order to fit a predictive model for calculation of CRC risk, an analysis of data from CRC patients and non-CRC patients is needed. Such a model could be used to assist physicians when making the decision whether an individual patient should have a colonoscopy.

5.2 Study aim and objectives

Inspired by the performance of the existing statistical models in calculating CRC risk, this study aimed to develop a predictive statistical model for calculation of CRC risk in the population of patients referred to the secondary care, based on symptoms and other patient information included in the free-text notes in electronic referrals to the District Health Board in Hamilton. To develop this model, this sub-study addressed the following objectives:

1. Extraction of symptoms, signs, family history of CRC, personal history of CRC or

polyps, test results, and comorbidities related to lower gastrointestinal tract from free-text notes using an automated procedure;

2. Development of a predictive model based on the extracted information and patients' demographics;

3. Internal validation of the developed model.

Additionally the study investigated the association between the time from first referral to colonoscopy and the following predictors: gender, age, ethnicity, and extracted variables.

5.3 Methods

5.3.1 Study design and study population

This was a retrospective cohort study conducted in Waikato region of New Zealand. The study population was all patients aged 18–90 years old, who were referred via e-referral to the Gastroenterology or General Surgery departments of the Waikato DHB hospital between 1 January 2015 and 31 December 2017, were clinically suspected of CRC or other lower gastrointestinal disease, and did not have preexisting CRC or polyps. Furthermore, the patients needed to have specified, in the free-text written by physicians in the referral letter, at least one of the following: a symptom or test result associated with CRC; family history of CRC; or a comorbidity related to the gastrointestinal tract.

The study cohort was derived from the study population. The reference standard used in this study was a complete colonoscopy (colonoscopy with visualisation of the caecum). Therefore only patients who had a full colonoscopy completed within the study period were included in the study cohort. In order to be included in the study, patients had

to have at least one e-referral made before their first colonoscopy. CRC patients who did not have at least one e-referral made before CRC diagnosis were also excluded from the analysis as, in order to study associations between symptoms and CRC risk, it was necessary to have indicated symptoms before CRC diagnosis. The outcome was diagnosis of CRC, defined as registration in NZCR from the date of the first colonoscopy after the referral, to the end of the study (31 December 2017).

5.3.2 Data

5.3.2.1 Data sources

Data used in this sub-study were obtained from three sources: electronic referrals (e-referrals) from the Midlands region in NZ to the Gastroenterology and General Surgery departments in the Waikato Hospital in Hamilton, between 1 January 2015 and 31 December 2017; a data set containing information about colonoscopies performed in the hospital in the years 2015–2017; and registrations with ICD-10 codes C18.0–C20.0 from the New Zealand Cancer Registry (NZCR) from 1January 2015 to 31 December 2017. The original data included patients' unique NHI numbers. In order to protect patients' privacy, an analyst from the Waikato DHB replaced the NHI numbers with encrypted unique patient identifiers which were consistent across the three data sets. For the analysis, I linked the data sets using the unique patient identifiers.

Briefly, among many variables in the three data sets, the following were used in the analysis:

1. from the e-referrals: demographics; date of referral; free-text notes, which were used to extract the information about symptoms, family history of CRC, laboratory test results, comorbidities, and information about polyps or CRC diagnosed prior to the referral;

2. from the colonoscopy data: the date of colonoscopy; and indication of full vs partial colonoscopy; and

3. from NZCR: date of diagnosis; tumour characteristics; and demographics.

5.3.2.2 Data preprocessing

The process of preparing the study cohort used for statistical analysis is presented in Figure 5.1. Before merging the three data sets, the NZCR data were reduced to one record per patient, as eight patients had two CRC diagnoses. Multiple rows for one patient were combined into one row in such a way that patients with colon and rectal cancers had both cancer diagnoses indicated, and the first date of diagnosis was used in analysis.

After e-referrals data were merged with the colonoscopy data and NZCR data, multiple colonoscopies were combined, and the date of the first colonoscopy was used in the analysis.

Free-text notes from all referrals prior to the first colonoscopy were retained and used for the extraction of symptoms and other information. At this stage the extraction of information from free-text notes for 23,686 patients was carried out, leading to the selection of the study population (patients without preexisting CRC or polyps, with at least one symptom relevant to lower gastrointestinal disease specified).

From the study population, the study cohort was selected based on the inclusion criteria, i.e., a patient needed to have a full colonoscopy, and at least one symptom (or other factor) specified preceding the colonoscopy. Also, the colonoscopy had to be performed prior to CRC diagnosis (if applicable).

The data set for the study cohort, constructed as explained above, contained a referral

(not a patient) as an observational unit, with some patients having multiple referrals. For the statistical analysis, however, each observational unit has to be a patient. In the case of multiple referrals, therefore, symptoms included in all referrals belonging to a particular patient were combined and included in the patient record so that all symptoms were still preserved despite, the reduction of the data set to the patient level.

5.3.2.3 Extraction of information from free-text notes

Symptoms consistent with CRC, test results, family history of CRC and comorbidities related to lower gastrointestinal disease as used in earlier studies (Ford et al., 2008; Jellema et al., 2010) were extracted from the free-text notes written by physicians and included in the e-referrals. The free-text notes were written for the purpose of patient care, not for scientific research. Therefore, there was no systematic coding of the symptoms, such as ICD-10 codes. All specified symptoms and other information about a patient gathered during a medical consultation were given in free-text format, which is problematic for searching, summarising and statistical analysis (Meystre et al., 2008).

I carried out the extraction of the information using an automated procedure, based on a spelling file made purposely for this study. Although there are available algorithms for extraction of information from unrestricted clinical language (for an overview see Ford et al. (2016)), I decided to develop my own script for the extraction of symptoms. This is because the already-available algorithms are designed mostly for a specific type of information, e.g., extraction of medication names (Xu et al., 2010), or do not have sufficient accuracy. For this sub-study a high accuracy of the symptom extraction was required, as the number of CRC patients was already low, so it was important to extract correct information for as many patients as possible. I aimed at 90% of symptoms being correctly extracted.

A list of words and phrases was prepared in order to identify symptoms described in

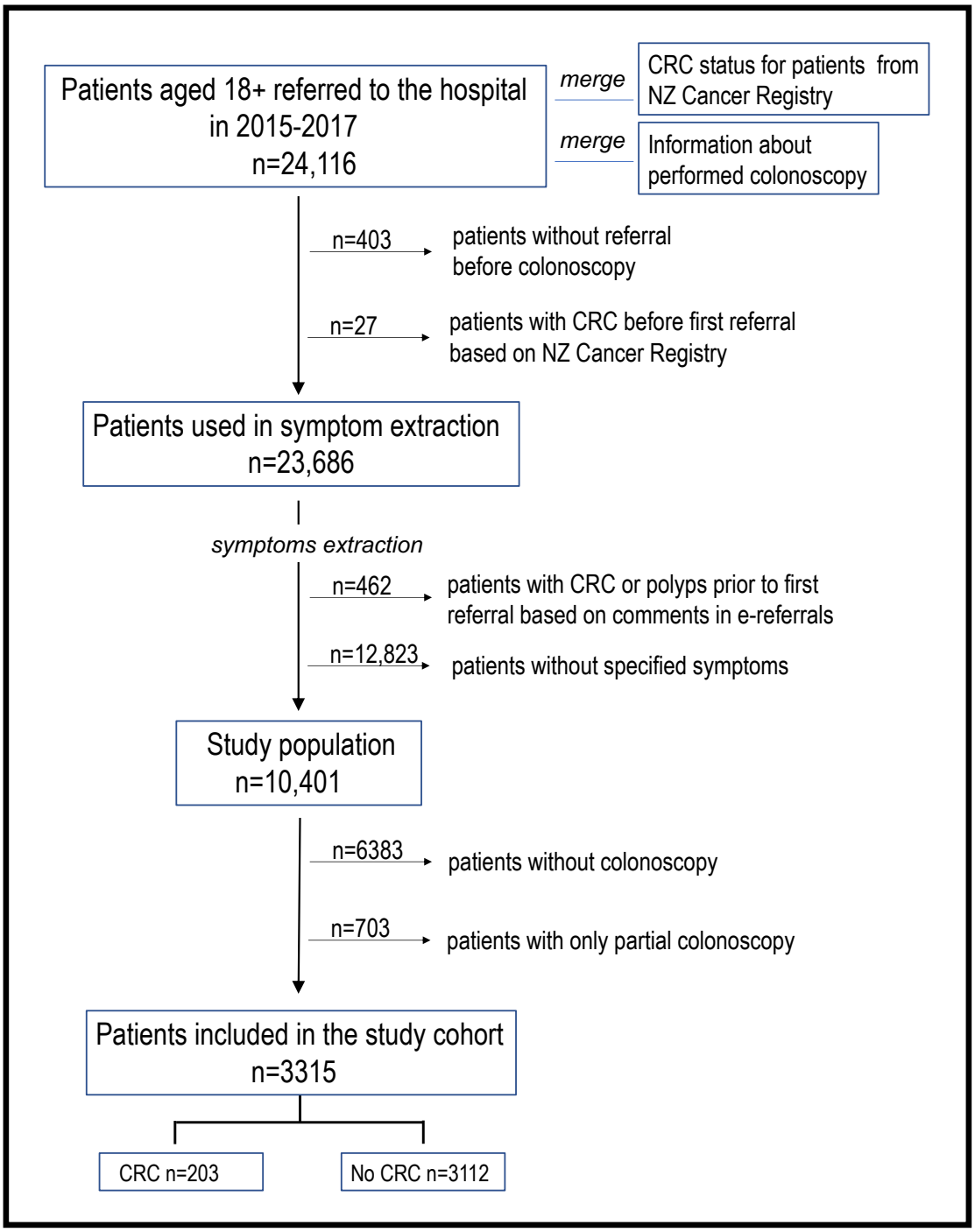


Figure 5.1: *Process of determination of the study cohort.*

comments for each patient, including different spelling variants for each key word. Only symptoms and other predictors indicated as present were extracted, which means that, e.g. "possible weight loss" was not extracted as weight loss. Key words used to extract

selected symptoms and other predictors are listed in Table 5.1.

Group of predictors	**Variable**	**Key words**
Symptoms and signs	abdominal pain	lower abdominal/epigastric/bowel/umbilical/ flank/iliac/fossa/hypochondrial pain/cramp/discomfort/tenderness
	bloating	bloating, abdominal distention, flatulence
	rectal bleeding	rectal/intestinal/ bleeding, blood in/on feaces/bowel motion/stool/poo/ toilet/paper, bloody/tinged/stained mucous, bright red/dark red/tarry/dark brown/ clothed/passed blood, hematochezia, melena
	constipation	constipation
	diarrhoea	diarrhoea, stools towards diarrhoea, explosive stools
	weight loss	any indicated weight loss
Test results	occult blood	occult blood, fit/FOBT positive
	anaemia	anaemia, iron deficiency/low level, low/dropping/declining/falling haemoglobin/ iron/ferritin
Family history		mother/father/uncle/aunt/siblings/ relatives had CRC, family history of CRC, familial
Comorbidities	inflammatory bowel disease	IBD, calprotectin rised, crohn, colitis, diverticulitis, proctitis
	haemorrhoids	haemorrhoids

Table 5.1: *Key words for extraction of information from free-text notes.*

Although care was taken to assure that all of the ways in which a particular symptom could be described were recognised, it is not possible to guarantee that all possible spelling

variations were included in the spelling file. The following quality control of the symptom extraction was carried out, therefore, in order to minimise errors in classification:

1. Free-text notes belonging to patients who were diagnosed with CRC were inspected manually for any inaccuracy in symptom extraction (a symptom not extracted but present, or symptom not present but extracted). Any unusual combination of words not specified in the spelling file and found, was corrected in the spelling file. In the first step, the subset of referrals belonging to CRC patients was used, as opposed to a random subset from the whole data set, because it is particularly important to classify all presented symptoms correctly in CRC patients. This is because the number of CRC patients is small compared to the entire data set, and any error would have a bigger influence on the results.

2. Under the assumption that there is no correlation between time of referral and symptoms, a second check was performed on the first 1000 referrals belonging to patients who received a colonoscopy

3. A threshold of 1.0% was used, allowing ten symptoms in the first 1000 referrals to be wrongly classified, as long as there was no systematic misclassification; that is, the misclassification of the same symptom did not happen more than twice. Each error in spelling of symptoms found during the checking process was corrected.

4. Finally, a random sample equal to 10% of all referrals used in symptom extraction was drawn and checked for the accuracy of the symptom extraction. Those 1000 referrals that were already checked in the previous step were excluded from the random draw.

5.3.3 Statistical analysis

This section starts with the explanation of the methods used for descriptive statistics, and subsequently explains steps undertaken in the development of the model for calculation of CRC risk in individual patients. The final part explains details of the additional investigation of factors that affected time from a referral to a colonoscopy. The effect of age at referral on the response variables might not be linear; therefore, for each of the two regression models (for CRC risk and for time to colonoscopy), I fitted two univariable regression models with age at referral on linear scale and on logarithmic scale as predictor. Subsequently, the deviances were compared in order to decide which representation of age to use in the multivariable models. Data were analysed using R version 3.2.2 software. A significance level of 5% was used throughout.

5.3.3.1 Descriptive statistics and univariable analysis

Data were summarised using mean and standard deviation for normally distributed variables, median and interquartile range for continuous but not normally distributed variables, and frequencies for categorical variables. Missing data in categorical variables were treated as a separate level. In order to investigate whether the study cohort were representative of the study population, for the key demographic variables, the study cohort was compared to the patients who satisfied inclusion criteria, but did not undergo the reference test (colonoscopy), and, for that reason, were excluded from the statistical analysis. Fisher's exact test was used to compare the distribution of categorical variables, and the Mann-Whitney test was used for comparison of the age distribution between two groups.

Two-by-two tables were constructed in order to compute: the crude association between given symptoms and CRC (the crude OR); the prevalence of symptoms; sensitivity; specificity; positive and negative predictive values (PPV and NPV); and positive and negative

likelihood ratios (PLR and NLR), all with corresponding 95% CI, using the R package epiR (Carstensen et al., 2019). The denominators for the calculation of diagnostic performance differed between symptoms. I used two different values for denominators: patients with unspecified abdominal symptoms were not included in the denominator for calculating values for the following symptoms: abdominal pain; bloating; and change in bowel habits. This is because the unspecified abdominal symptoms could include: abdominal pain; bloating; or change in bowel habits; and determination of the number of patients for nominator was impossible. However, patients with unspecified abdominal symptoms were included in the denominator for the calculation of diagnostic performance values for the remaining symptoms and other predictors.

Prevalence of symptoms was explored in relation to ethnicity and gender, and with respect to the anatomical location of tumours (proximal colon/distal colon/rectum). The frequencies of symptoms were compared between the investigated groups of patients using a Chi-square test. In addition, the age distribution for the most prevalent symptoms was also presented. The median time from first referral to colonoscopy was compared between males and females, and between Māori and non-Māori, using a Mann-Whitney test.

5.3.3.2 Statistical model for CRC risk

To develop a predictive model for assessing CRC risk in patients referred to secondary care, I carried out multivariable logistic regression analysis with CRC status (yes/no) as a response variable. All extracted symptoms, test results, family history of CRC and comorbidities, along with patients' age (modelled on logarithmic scale), gender and prioritised ethnicity (categorised as Māori/ non-Māori/ unknown), were included as predictors.

I considered inclusion of interaction terms in the model. Interaction terms were selected

using regression tree analysis (Camp and Slattery, 2002). Five regression trees were fitted using the rpart package (Therneau et al., 1997) with the complexity parameter (defined as the minimal R^2 improvement per split) set to 0.0001, and the minimal number of individuals per terminal node set to 30. The control parameters were adapted from Buchner et al. (2017). The five fitted trees differed with respect to the level of pruning, controlled by the *nsplit* parameter, as explained in Hothorn and Everitt (2014, Chapter 8). The final tree was chosen on the basis of the cross-validated area under the receiver operating characteristic curve (AUC). For fitting the regression tree, I used the same variables which were predictors in the logistic model. All two-way interaction terms detected by the final tree were chosen for inclusion in the logistic regression model (see the tree in the result section in Figure 5.7). In the next step, two following multivariable logistic regression models were fitted. First, a model which included only the following main terms: all extracted symptoms; test results; family history of CRC; comorbidities; age; and gender. And, second, a model which included all the main terms, and all two-way interactions suggested by the selected regression tree.

Both initial models were simplified by applying backwards elimination using AIC as the selection criterion for main terms. Interaction terms were kept in the model if they were statistically significant ($p<0.05$). Each of the simplified models was internally validated using stratified cross-validation, as recommended by Smith et al. (2014). Each of 203 folds contained one CRC case, and the controls were randomly split between folds. For both models, the ROC curves were constructed, and the AUC was calculated for the two models and compared (Pepe et al., 2003). The model with the higher AUC was chosen as the final predictive model. The final model's fit to the data was assessed using the Hosmer-Lemeshow test (Hosmer Jr et al., 2013).

Subsequently, the chosen model was corrected for optimism by shrinkage of the regression coefficients towards zero. This method is one way of improving predictions from a regression model in future patients. I applied uniform shrinkage with the shrinkage

factor determined using bootstrapping, based on Steyerberg and Vergouwe (2014). 1000 random re-samples of the data were used to fit the model, and the obtained values of regression coefficients (1000 sets of coefficients) were used to construct the shrinkage factor. The shrinkage factor was subsequently applied to the coefficients in the chosen model to give the final model. The model with shrunken coefficients will give more accurate predictions of the risk of CRC in the future secondary care population, compared to the initial model, which is likely to represent an overfit. To assess the performance of the final model in practice, the CRC risk for each patient was calculated using cross validation as explained above. Subsequently, I explored the consequences of performing colonoscopy only on those patients who had the highest CRC risk (according to the final model); the proportion of patients that would still be diagnosed with CRC if the number of performed colonoscopies was reduced by 20% and 40% was reported, following the methodology from Adelstein et al. (2010).

5.3.3.3 Analysis of time to colonoscopy

The differences in waiting time from referral to colonoscopy between Māori and non-Māori, and between males and females were assessed using Kaplan-Meier curves and log-rank tests. To explore whether any particular symptoms influence time to colonoscopy, and whether the time to colonoscopy depends on age, gender or ethnicity, a multivariable Cox regression model was fitted, with time from the first referral to colonoscopy as the response variable, and all symptoms, test results, comorbidities, age at referral, gender and prioritised ethnicity (Māori/ non-Māori/ unknown) as predictors. Backwards elimination using AIC as a criterion for selection of variables for the final model terms was applied to simplify the model.

5.4 Results

5.4.1 Study population and study cohort

There were 24,116 patients aged 18+ years during the study period who had a referral to the hospital in Hamilton. Of those, 403 patients had not had a referral before colonoscopy (they must have had a referral before the study period) and were removed from the data set. From 1 January 2015 to 31 December 2017, there was 503 new registrations of CRCs according to NZCR, with eight patients having two CRC diagnoses, giving 495 patients newly diagnosed with CRC. Patients who had a CRC diagnosis, according to NZCR, before their first referral (n=27) were removed, leaving 23,686 patients for extraction of information from free-text notes. Based on extracted information, 462 patients were removed from the study because they had a CRC or polyps diagnosis before the study period, according to information written in the free-text by GPs. Subsequently, 12,823 patients without any symptom, sign or other information relevant to lower gastrointestinal disease were removed from the data set. The remaining 10,401 patients constituted the study population.

To derive the study cohort (patients for statistical analysis), patients who did not undergo a full colonoscopy (reference test) were removed (n=7086). The remaining 3315 patients who had had full colonoscopy with visualisation of cecal pool, and fulfilled the inclusion criteria, constituted the study cohort. Within this cohort, there were 203 patients newly diagnosed with CRC during the study period (Figure 5.1).

In the following parts of this sub-study, the term *all patients* refers to the study cohort. Results and discussion related to any subgroup will be indicated in the text. The term '*age* refers to patients' age at referral.

5.4.2 Patients and tumour characteristics

Among 3315 patients included in the cohort, 42% were males. The age range in the cohort was 18–94 years, with a median age of 64 years. The median age in Māori patients (58.5 years) was lower than in non-Māori (66.0 years), which reflects the difference in age distribution between Māori and non-Māori. There were 7086 patients who had specified symptoms and fulfilled the inclusion criteria, but who did not have a full colonoscopy performed to determine the outcome, and therefore were not included in statistical analysis. Demographics of the study cohort are presented in Table 5.2. Restricting the cohort used in the statistical analysis to patients who underwent a full colonoscopy, affected the representativeness of the study cohort with respect to age and ethnicity, but not with respect to gender. Māori and younger patients were underrepresented in the study sample ($p<0.001$ for both tests).

	Study cohort (n=3315)			**Patients without full colonoscopy (n=7086)**
X	**All**	**Māori**	**non-Māori**	
Gender				
Female	1932	204	1690	4095
Male (%)	1383 (42)	138(40)	1128(40)	2991 (42)
Age				
Median	64	58.5	66	57
(IQR)	(52; 72)	(49; 66)	(53; 73)	(41; 71)
Ethnicity				
Māori (%)	342 (10)	-	-	1042 (15)
Non-Māori (%)	2818 (85)	-	-	5760 (81)
Not known (%)	155 (5)	-	-	284 (4)

Table 5.2: *Demographics for the study cohort and for patients who had specified symptoms in their referral but did not have a full colonoscopy.*

The prevalence of CRC in the study cohort was 6.1%. The median age of the 203 CRC patients was 70 years. In Māori patients, the median age was 64 years. Among all CRC patients, 32% of patients were diagnosed in early stage I, compared to 12% in distant metastatic stage (stage IV). The stage of CRC at diagnosis was not known in 27% of patients (n=54). It is quite likely that many of the unknown stages are due to diagnosis in very advanced stages, which shifts the proportion in favour of early stages. Sixty-six percent of tumours were located in the colon. Four patients were diagnosed with both colon and rectal cancer at the same time. Characteristics of patients diagnosed with CRC, and the frequencies for stages of disease at diagnosis, and for the tumour site, are presented in Table 5.3. The frequencies are given for the whole group of CRC patients, for Māori and non-Māori.

Māori were not compared to non-Māori with respect to characteristics of CRC because of the small number of Māori patients diagnosed with CRC in this study cohort (n=22). However, based on the entire data set containing new CRC diagnoses registered on the NZCR (n=503), of whom 54 (10.7%) were Māori, Māori were more often diagnosed with CRC in metastatic stage IV than non-Māori (48% versus 22%), and the difference was highly statistically significant ($p<0.001$, Fisher's test). With respect to tumour location, tumours of the colon were present more often in non-Māori (71%) than in Māori (57%).

Variable	All	Māori	non-Māori
	n=203	n=22	n=169
Gender			
Female	94	11	79
Male(%)	109 (54)	11 (50)	90 (54)
Age at referral			
Median	70	64	71
(IQR)	(63.5; 79)	(57; 69)	(65; 80)
Stage at diagnosis			
I	65 (32)	6 (27)	57 (34)
II	17 (8)	1 (5)	13 (8)
III	43 (21)	5 (23)	36 (21)
IV	25 (12)	2 (9)	21 (12)
Not known	54 (27)	8 (36)	42 (25)
Tumour site			
colon	135 (66)	13 (59)	115 (68)
rectum	64 (32)	9 (41)	50 (30)
colon and rectum	4 (2)	0	4 (2)

Table 5.3: *Characteristics of patients with CRC (n=203) with respect to demographics and disease description with respect to stage of CRC at diagnosis and site of the tumour (colon/rectum). The frequencies for Māori and non-Māori do not add up because frequencies for patients with unknown ethnicity are included in the total.*

5.4.3 Symptoms, test results, family history of CRC and comorbidities

Table 5.4 presents fourteen symptoms related to lower gastrointestinal disease, as well as the following additional variables: occult blood; anaemia; family history of CRC; and comorbidities related to the gastrointestinal tract. The table presents comprehensive information for the all extracted variables, including: the frequency; prevalence; values of

diagnostic performance; specificity; sensitivity; positive predictive value (PPV); negative predictive value (NPV); positive likelihood ratio (PLR); negative likelihood ratio (NLR); and crude ORs, all with 95% CIs. All symptoms had high (93–96%) NPVs for CRC, reflecting the low prevalence of CRC in the study cohort. Positive predictive values for CRC varied considerably, with the highest PPVs for palpable mass in abdomen or rectum (PPV= 16%), anaemia (PPV= 14%), and for rectal bleeding (PPV= 12%), and with the lowest PPV= 1% for IBD and 2% for family history of CRC.

Symptom	TP	FP	FN	TN	prevalence	sensitivity	specificity	PPV	NPV	PLR	NLR	OR
Abdominal pain	19	610	175	2468	0.19	0.10	0.80	0.03	0.93	0.49	1.13	0.44
					(0.18;0.21)	(0.06;0.15)	(0.79;0.82)	(0.02;0.05)	(0.92;0.94)	(0.32;0.76)	(1.07;1.2)	(0.27;0.71)
Anal symptoms	3	44	200	3068	0.01	0.01	0.99	0.06	0.94	1.05	1.00	1.05
					(0.01;0.02)	(0.00;0.04)	(0.98;0.99)	(0.01;0.18)	(0.93;0.95)	(0.33;3.34)	(0.98;1.0)	(0.32;3.40)
Bloat	2	58	192	3023	0.02	0.01	0.98	0.03	0.94	0.55	1.01	0.54
					(0.01;0.02)	(0.00;0.04)	(0.98;0.99)	(0.00;0.12)	(0.93;0.95)	(0.13;2.23)	(0.99;1.0)	(0.13;2.24)
Misc symptoms	4	80	199	3032	0.03	0.02	0.97	0.05	0.94	0.77	1.01	0.76
					(0.02;0.03)	(0.01;0.05)	(0.97;0.98)	(0.01;0.12)	(0.93;0.95)	(0.28;2.07)	(0.99;1.0)	(0.28;2.10)
Change in bowel habit	57	1046	137	2035	0.34	0.29	0.66	0.05	0.94	0.87	1.07	0.81
					(0.32;0.35)	(0.23;0.36)	(0.64;0.68)	(0.04;0.07)	(0.93;0.95)	(0.69;1.08)	(0.97;1.2)	(0.59;1.11)
Constipation	8	115	195	2997	0.04	0.04	0.96	0.07	0.94	1.07	1.00	1.07
					(0.03;0.04)	(0.02;0.08)	(0.96;0.97)	(0.03;0.12)	(0.93;0.95)	(0.53;2.15)	(0.97;1.0)	(0.51;2.22)
Dark blood	9	66	194	3046	0.02	0.04	0.98	0.12	0.94	2.09	0.98	2.14
					(0.02;0.03)	(0.02;0.08)	(0.97;0.98)	(0.06;0.22)	(0.93;0.95)	(1.06;4.13)	(0.95;1.0)	(1.05;4.36)
Diarrhoea	8	275	195	2837	0.09	0.04	0.91	0.03	0.94	0.45	1.05	0.42
					(0.08;0.10)	(0.02;0.08)	(0.90;0.92)	(0.01;0.05)	(0.93;0.94)	(0.22;0.89)	(1.02;1.1)	(0.21;0.87)
Lack of appetite	1	23	202	3089	0.01	0.00	0.99	0.04	0.94	0.67	1.00	0.66
					(0.00;0.01)	(0.00;0.03)	(0.99;1.00)	(0.00;0.21)	(0.93;0.95)	(0.09;4.91)	(0.99;1.0)	(0.09;4.95)
Palpable mass	8	42	195	3070	0.02	0.04	0.99	0.16	0.94	2.92	0.97	3.00
					(0.01;0.02)	(0.02;0.08)	(0.98;0.99)	(0.07;0.29)	(0.93;0.95)	(1.39;6.14)	(0.95;1.0)	(1.39;6.48)
Mucous	3	43	200	3069	0.01	0.01	0.99	0.07	0.94	1.07	1.00	1.07
					(0.01;0.02)	(0.00;0.04)	(0.98;0.99)	(0.01;0.18)	(0.93;0.95)	(0.33;3.42)	(0.98;1.0)	(0.33;3.48)
Tiredness	4	47	199	3065	0.02	0.02	0.98	0.08	0.94	1.30	1.00	1.31
					(0.01;0.02)	(0.01;0.05)	(0.98;0.99)	(0.02;0.19)	(0.93;0.95)	(0.47;3.59)	(0.98;1.0)	(0.47;3.67)
Weight loss	28	288	175	2824	0.10	0.14	0.91	0.09	0.94	1.49	0.95	1.57
					(0.09;0.11)	(0.09;0.19)	(0.90;0.92)	(0.06;0.13)	(0.93;0.95)	(1.04;2.14)	(0.90;1.0)	(1.03;2.38)
Rectal bleeding	73	916	130	2196	0.30	0.36	0.71	0.07	0.94	1.22	0.91	1.35
					(0.28;0.31)	(0.30;0.43)	(0.69;0.72)	(0.06;0.09)	(0.93;0.95)	(1.01;1.48)	(0.82;1.0)	(1.00;1.81)

Test result	TP	FP	FN	TN	prevalence	sensitivity	specificity	PPV	NPV	PLR	NLR	OR
Anaemia	81	499	122	2613	0.17	0.40	0.84	0.14	0.96	2.49	0.72	3.48
					(0.16;0.19)	(0.33;0.47)	(0.83;0.85)	(0.11;0.17)	(0.95;0.96)	(2.06;3.00)	(0.64;0.8)	(2.58;4.68)
Occult blood	16	165	187	2947	0.05	0.08	0.95	0.09	0.94	1.49	0.97	1.53
					(0.05;0.06)	(0.05;0.12)	(0.94;0.95)	(0.05;0.14)	(0.93;0.95)	(0.91;2.43)	(0.93;1.0)	(0.90;2.61)
Family history	5	294	198	2818	0.09	0.02	0.91	0.02	0.93	0.26	1.08	0.24
					(0.08;0.10)	(0.01;0.06)	(0.89;0.92)	(0.01;0.04)	(0.92;0.94)	(0.11;0.62)	(1.05;1.1)	(0.10;0.59)
Comorbidities												
IBD	3	307	200	2805	0.09	0.01	0.90	0.01	0.93	0.15	1.09	0.14
					(0.08;0.10)	(0.00;0.04)	(0.89;0.91)	(0.00;0.03)	(0.92;0.94)	(0.05;0.46)	(1.07;1.1)	(0.04;0.43)
Haemorrhoids	9	141	194	2971	0.05	0.04	0.95	0.06	0.94	0.98	1.00	0.98
					(0.04;0.05)	(0.02;0.08)	(0.95;0.96)	(0.03;0.11)	(0.93;0.95)	(0.51;1.89)	(0.97;1.0)	(0.49;1.95)
Abn liver function	3	32	200	3080	0.01	0.01	0.99	0.09	0.94	1.44	1.00	1.44
					(0.01;0.01)	(0.00;0.04)	(0.99;0.99)	(0.02;0.23)	(0.93;0.95)	(0.44;4.65)	(0.98;1.0)	(0.44;4.76)

Table 5.4: *Diagnostic performance of symptoms and other predictors extracted from free-text notes*

Less than half of the cohort presented with more than one symptom, with 29 patients having specified five or more symptoms. Among CRC patients, 50% presented with more than one symptom. The frequencies are presented in Table 5.5.

Number of symptoms	1	2	3	4	5 +
Number of patients (n=3315)	1800	1032	355	99	29
Number of patients with CRC (n=203)	101	68	23	9	2
Percent CRC	5.5	6.5	6.5	9.1	6.9

Table 5.5: *Number of extracted symptoms per patient.*

In the study cohort, the most often reported symptoms were: change in bowel habit (33%); rectal bleeding (30%); abdominal pain (19%); anaemia (18%); and weight loss (10%). The most often reported combination of two symptoms was change in bowel habit accompanied by rectal bleeding, reported in 209 patients. The frequencies of the most often reported pairs of symptoms in are presented in Table 5.6.

Symptom (n with the symptom)	**Accompanying symptoms**	**Number of patients with the combination**
Change in bowel habit	Rectal bleeding	209
(n=1103)	Abdominal pain	186
	Weight loss	137
	Anaemia	95
Rectal bleeding	Abdominal pain	117
(n=990)	Anaemia	102
Abdominal pain	Weight loss	88
(n=629)		

Table 5.6: *Combination of two symptoms occurring the most frequently.*

The distribution of age differed between symptoms and other predictors. Patients presenting with a family history of CRC and IBD had the lowest median age. The symptoms anaemia, occult blood, dark blood, palpable mass and weight loss were more common in older patients. The age distributions for the most frequent and important symptoms are presented in Figure 5.2.

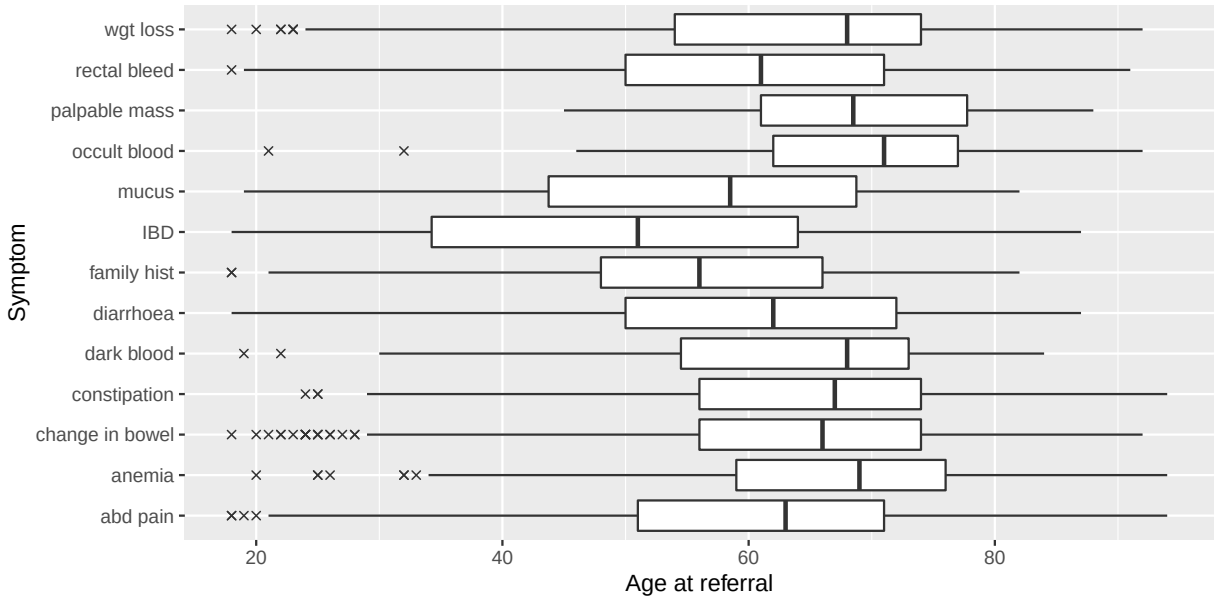


Figure 5.2: *Distribution of age at referral by symptom in the study cohort.*

With respect to stage of disease, palpable mass was present nearly exclusively in combined late stages III and IV (D+E), while constipation was nearly exclusively reported in patients diagnosed in combined early stages I and II (B+C) (Figure 5.3).

With respect to tumour site (colon/rectum), anaemia was more prevalent in patients with cancer of the colon, while change in bowel habits and rectal bleeding were more prevalent in patients with rectal cancer (Figure 5.4). The frequencies of symptoms by stage and by anatomical site are also given in Table 5.7.

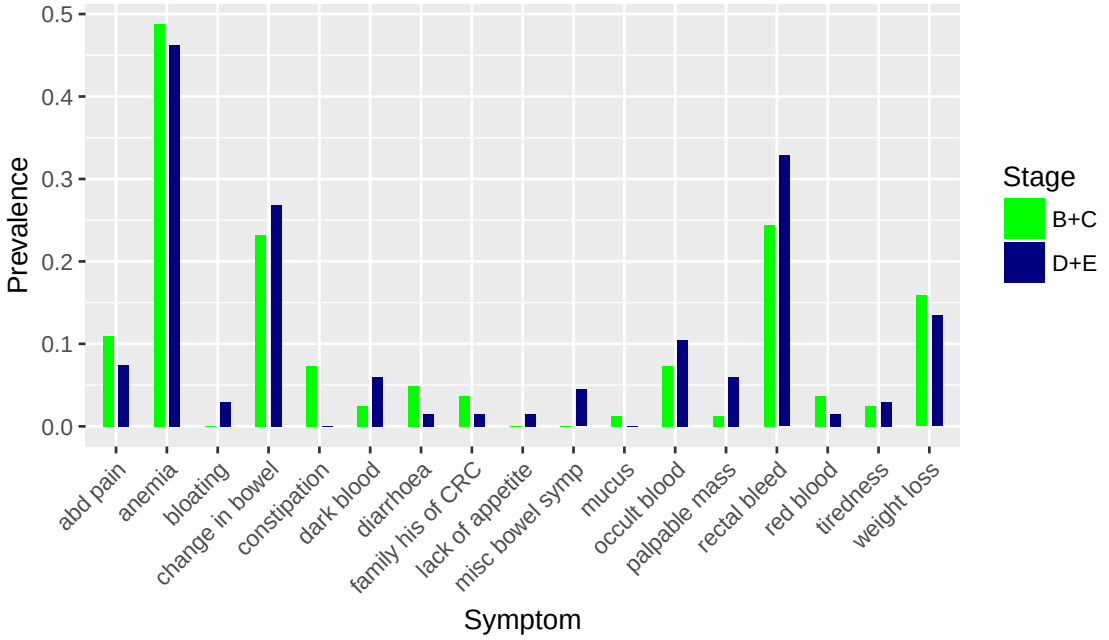


Figure 5.3: *Prevalence of symptoms and test results in combined early I+II, and late III+IV stages (n=149). Patients with unknown stage of CRC were excluded from this analysis.*

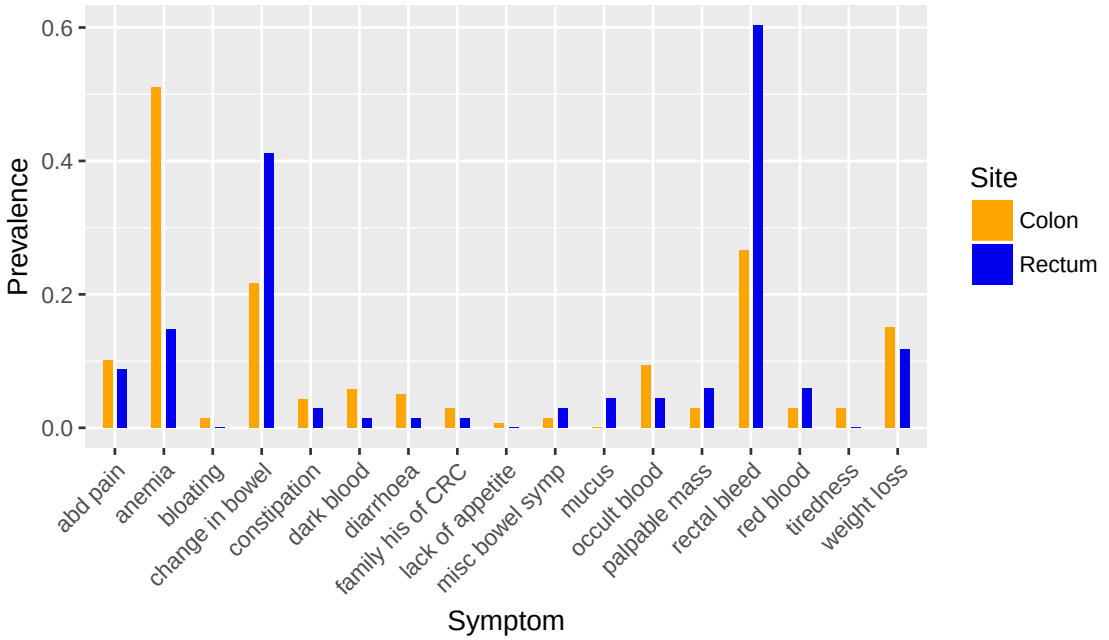


Figure 5.4: *Prevalence of symptoms and test results by site of tumour: colon or rectum (n=203).*

Symptom	I+II	III+IV	Colon	Rectum
Abdominal pain	9	5	14	6
Anaemia	40	31	71	10
Bloating	0	2	2	0
Change in bowel habits	19	18	30	28
Constipation	6	0	6	2
Dark blood	2	4	8	1
Diarrhoea	4	1	7	1
Family history	3	1	4	1
IBD	1	0	2	1
Lack of appetite	0	1	1	0
Misc symptoms	0	3	2	2
Mucous	1	0	0	3
Occult blood	6	7	13	3
Palpable mass	1	4	4	4
Rectal bleeding	20	22	37	41
Red blood	3	1	4	4
Tiredness	2	2	4	0
Weight loss	13	9	21	8

Table 5.7: *Frequencies of symptoms and test results for combined stages of CRC and for location of tumours: colon/rectum.*

5.4.4 Time to colonoscopy

The median time from the first referral to the first colonoscopy for the whole cohort was 115 (IQR; 49, 191) days. In Māori patients, the median waiting time was longer than in non-Māori - 132 days versus 112 days (p=0.037, logrank test). There was no statistically significant difference between time to colonoscopy in males and females (112 vs 117 days respectively, p=0.057 for logrank test). The distributions of the time to colonoscopy by ethnicity and gender are shown as Kaplan-Meier curves in Figure 5.5.

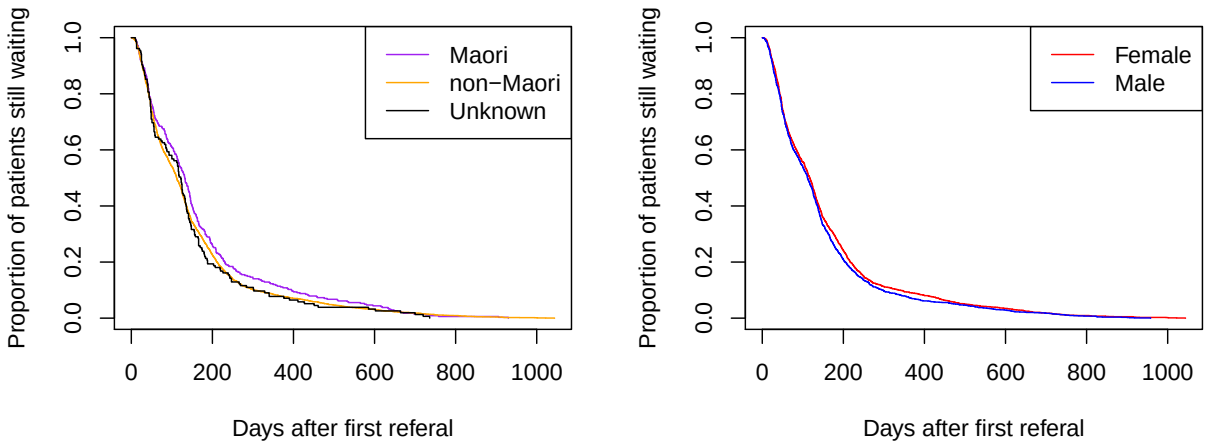


Figure 5.5: *Kaplan-Meier curves for time from first referral to colonoscopy by ethnicity and gender. Only patients who had full colonoscopy are included.*

However, in the final Cox model adjusted for age and other predictors, ethnicity was not included as it was removed by the backwards elimination procedure. Gender was also not associated with waiting time. The following symptoms were found to be statistically significantly associated with a shorter time from referral to colonoscopy: anaemia, occult blood and rectal bleeding, with anaemia and occult blood having the highest influence on shortening the time to colonoscopy. The symptoms which increased the time to colonoscopy were abdominal pain, abnormal liver function and haemorrhoids. The exponentiated coefficients (HRs) with 95% CIs are shown in Table 5.8. For visual comparison, the HRs for the symptoms included in the final age-adjusted Cox model are shown in Figure 5.6.

		95% CI	
X	**OR**	**Lower**	**Upper**
abdominal pain	0.79	0.72	0.87
anaemia	1.32	1.20	1.45
anal symptoms	0.80	0.60	1.07
haemorrhoids	0.71	0.60	0.84
impaired liver functions	0.70	0.50	0.98
rectal bleeding	1.14	1.06	1.24
weight loss	1.10	0.98	1.23
age at referral	1.005	1.002	1.007

Table 5.8: *Exponentiated coefficients (HRs) from the final model for time from first referral to colonoscopy.*

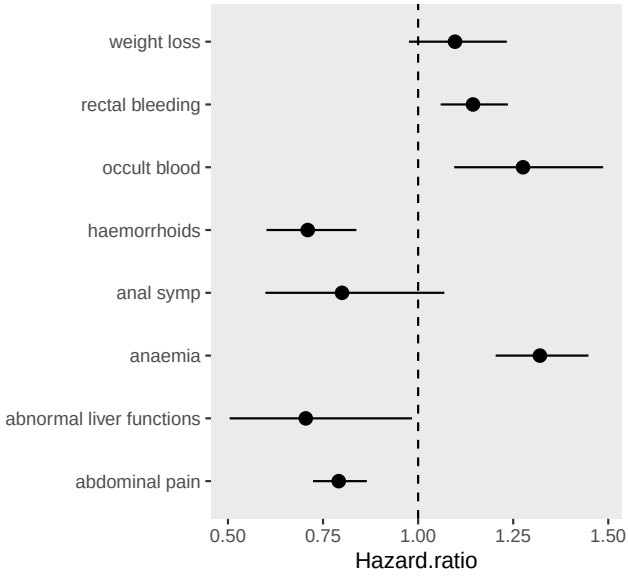


Figure 5.6: *Values of the coefficients with 95% CIs from age-adjusted Cox regression model for the associations between symptoms and the time from first referral to colonoscopy. The dashed vertical line represents no effect of a symptom on time from first referral to colonoscopy.*

5.4.5 Predictive model for CRC risk

Age at diagnosis was log transformed. For the purpose of the predictive model, ethnicity had three levels (Māori/non-Māori/not known). Patients with 'not known' ethnicity were included in the analysis in order to analyse all data available. After backward elimination, Model 2 (without interaction terms) included 9 predictors (the predictors are listed in Table 5.9). The area under the ROC curve for the cross-validated model was 0.77, compared to 0.79 for the non-cross-validated model. The reduction in AUC represents the reduction in the discriminatory power of the model, which we would expect when applying the model to future patients.

Model	**Terms in the reduced model**	**AUC (SE)**	
		not x-val*	**(x-val)**
1.**Regression tree**	age:anaemia, age:mass, age:rectal bleed, age:weight loss, anaemia:weight loss	0.77 (0.02)	0.69 (0.02)
2.**Model without interactions**	abdominal pain, anaemia, IBD, family history, lack of appetite, mass, rectal bleed, weight loss, gender, age	0.77 (0.02)	0.76 (0.02)
3.**Model with interactions**	main terms: abdominal pain, anaemia, diarrhoea, family history, IBD, lack of appetite, mass, rectal bleed, weight loss, age, gender interaction terms: anaemia:gender	0.78 (0.02)	0.76 (0.02)

*x-val: cross-validated

Table 5.9: *Fitted models*

Five regression trees, corresponding to different values of the *nsplit* parameter (8, 10, 12, 14 and 16), pruned to a maximum depth of 4, were fitted, and the tree with the highest

AUC after cross-validation (AUC=0.69) was chosen for identification of interaction terms. The tree is presented in Figure 5.7, and the identified interactions are listed in Table 5.9.

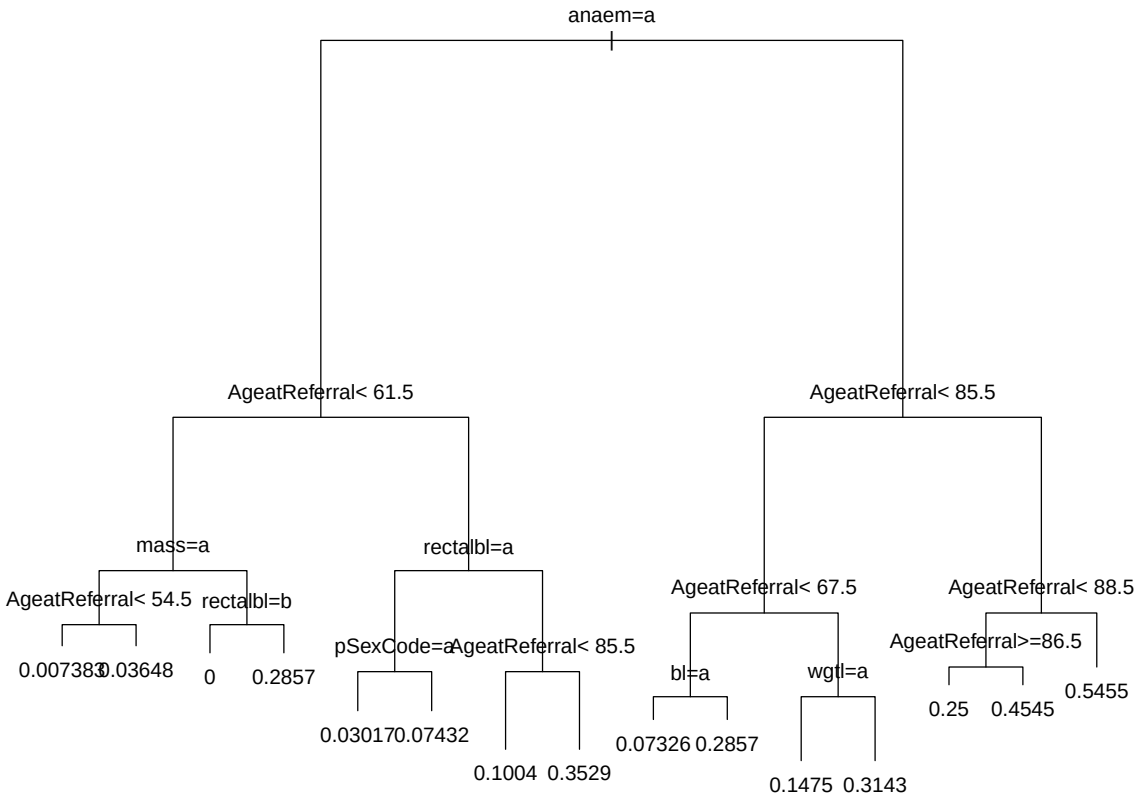


Figure 5.7: *Tree used for identification of interaction terms.*

After backwards elimination of Model 3, which initially included all main terms and all interaction terms identified by the regression tree, seven main terms and one interaction term were left. The AUC for the cross-validated model was 0.77, nearly the same as for Model 2. Table 5.9 summarises the AUC values and variables in the reduced models. The model without interactions was chosen as the preferred model, because the addition of interaction terms did not improve the predictive accuracy, judged by the value of AUC, in the cross-validated models.

The bootstrap procedure applied to the chosen model gave the shrinkage factor equal to 0.914. One symptom, namely lack of appetite, had a huge bias (-4.537) compared to the other predictors. This suggests that the value of the shrinkage factor is strongly influenced by lack of appetite. The values of biases on the coefficient estimates for the model, including lack of appetite, are given in Table 5.10.

Variable	**Bias**	
	Model with lack of appetite	**Model without lack of appetite**
intercept	-0.133	-0.217
abdominal pain	-0.032	-0.031
anaemia	-0.002	0.000
diarrhoea	-0.089	–
family history	-0.170	-0.168
IBD	-0.819	-0.816
lack of appetite	-4.537	–
mass	-0.078	-0.063
weight loss	-0.006	-0.007
rectal bleeding	-0.001	-0.001
gender	-0.006	-0.007
log(age)	0.030	0.029
AUC	-0.008	-0.005
shrinkage factor	0.914	0.934

Table 5.10: *Presentation of the bias in model with lack of appetite and in the model without lack of appetite.*

The model without "lack of appetite" was chosen as the final model. After lack of appetite was removed from the chosen model, the backwards elimination, bootstrapping,

coefficient shrinking, and the cross-validation procedures were repeated. The regression coefficients with 95% CI given by the final model with shrunken coefficients are presented in Table 5.11 and in Figure 5.8.

		95% CI	
X	**OR**	**Lower**	**Upper**
abdominal pain	0.56	0.34	0.92
anaemia	2.64	1.91	3.64
family history	0.46	0.19	1.16
IBD	0.29	0.09	0.94
palpable mass	2.55	1.13	5.75
rectal bleeding	1.71	1.24	2.37
weight loss	1.46	0.94	2.27
gender	1.57	1.17	2.11
log(age)	22.26	9.14	54.20

Table 5.11: *Exponentiated coefficients (ORs) from the final model for CRC risk with shrunk coefficients. The intercept is 8.94e-08 (95%CIs 1.78e-09, 3.47e-06).*

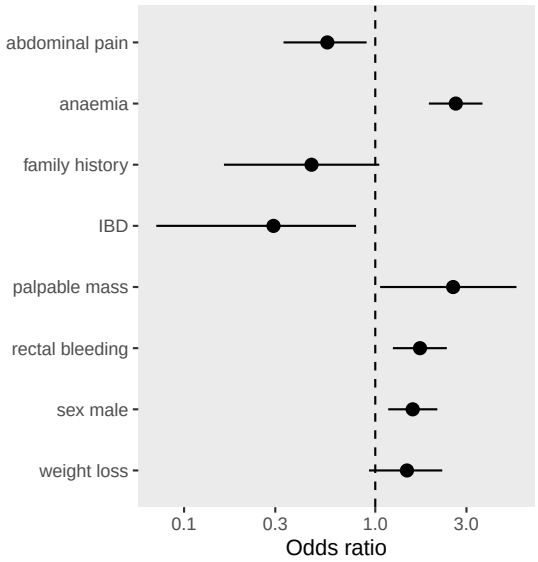


Figure 5.8: *Exponentiated coefficients (ORs) from the final model for CRC risk with 95% CIs (age adjusted). The dashed line represents no effect.*

Based on the final model, the CRC risk for patients presenting with combinations of symptoms corresponding to a high risk scenario (rectal bleeding and anaemia) and to a low risk scenario (abdominal pain and family history of CRC), separate for males and for females, are shown in Figure 5.9. As can be seen, the difference between CRC risk under those two scenarios is very big.

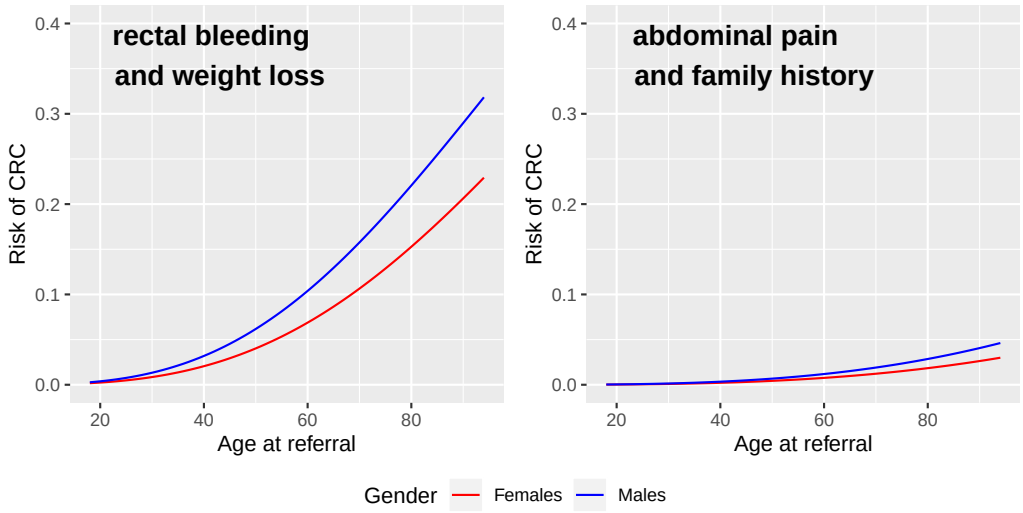


Figure 5.9: *Model based risk of CRC in patients with selected combinations of risk factors.*

According to the final model (Table 5.11), if a colonoscopy had been performed on only the 60% of patients from our cohort with the highest risk of CRC, only 89% of CRCs would be detected, and 23 patients out of 203 would not have been diagnosed. If a colonoscopy had been performed on 70% of patients, 96% of CRCs would be detected, and 9 patients would not have been diagnosed. To detect over 98% of cancers, it would be enough to perform 80% of the colonoscopies. In this case, in our cohort, four patients out of 203 would not have been diagnosed.

Figure 5.10 shows the ROC curve for the final cross-validated model with the AUC=0.76 (95% CI; 0.73, 0.79). The dots represent individual predictors. All predictors had high specificity and low sensitivity.

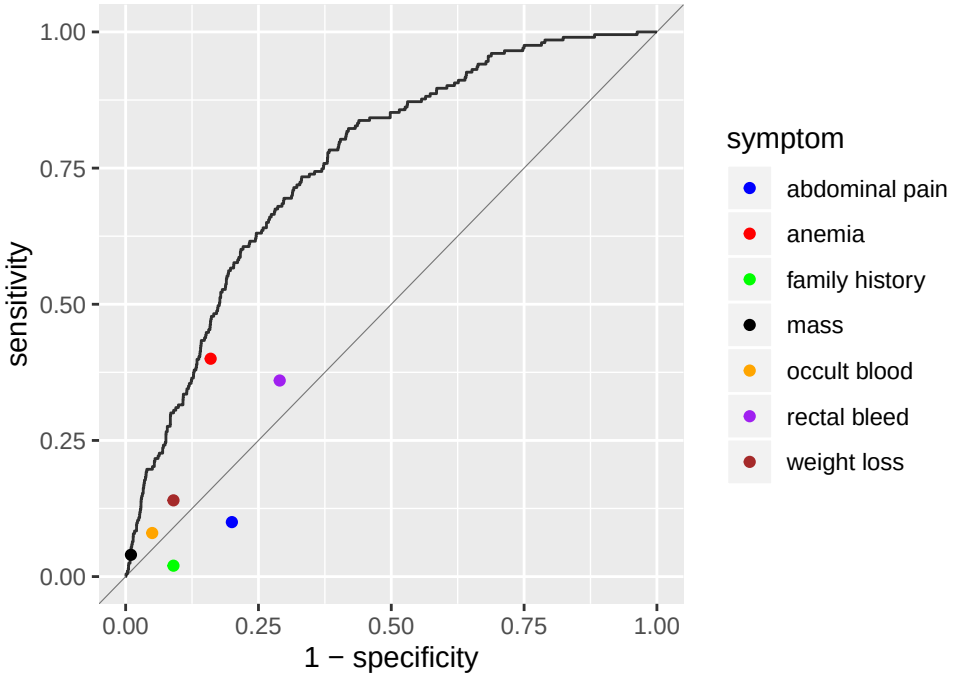


Figure 5.10: *ROC curve from the final model compared to the sensitivities and specificities for individual symptoms.*

There was no disagreement between the observed and fitted values (p=0.35 for the Hosmer-Lemeshow test). The graphical illustration of the model fit is presented in Figure 5.11.

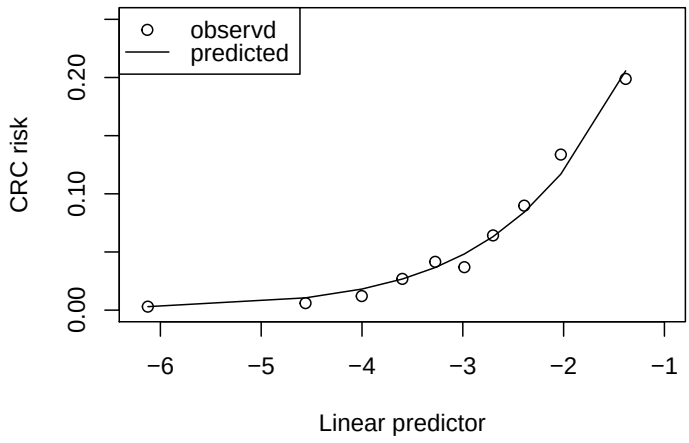


Figure 5.11: *Goodness-of-fit for the final model for CRC risk.*

5.5 Discussion

This section starts with a summary of findings, combined with a discussion and comparison to results from other studies. The results of this sub-study are comparable only to the results of studies that analysed data from the referred population, that is, a population with a much higher CRC risk and a higher prevalence of symptoms related to CRC than in the general or primary care population. The comparison therefore will be limited to studies which analysed data collected from secondary care patients. Further, the limitations and strengths are provided to guide the explanation of the possible use of the developed model in clinical practice. Finally, I provide a direction for the future research which emerged from this sub-study.

5.5.1 Summary and discussion of main findings

Symptoms which were used in this sub-study to develop the statistical model for CRC risk had, in general, lower prevalences in my cohort than in other studies (Selvachandran et al., 2002; Bjerregaard et al., 2007; Thompson et al., 2008; Adelstein et al., 2010). Those studies elicited symptoms present in patients using a questionnaire administered to patients. As a result, some symptoms which were associated with CRC in this sub-study - abdominal pain; change in bowel habit; rectal bleeding; and weight loss - had a relatively low prevalence compared to the studies mentioned. This is not surprising as, when filling out a questionnaire, patients will be prompted to think about those symptoms, while they may forget to mention them to a GP, or they might be embarrassed to talk about some symptoms. A recent study from NZ reported that 12% of their responders felt embarrassed about reporting some bowel symptoms to a GP, including altered bowel habit or rectal bleeding (Windner et al., 2018). By contrast, anaemia had a relatively high prevalence in this sub-study, compared to other studies. This can be explained by GPs having access to the laboratory results, while patients might not always be aware of

the result of the test for anaemia. Family history had a similar prevalence, which is not surprising as patients will nearly always be aware of it and, if so, they will also inform their GPs.

When comparing my results to studies which analysed data from the NZ population, a similar pattern can be observed. Hsiang et al. (2013), who analysed data from pre-diagnosis case notes from Auckland, reported the sensitivities for the most commonly reported symptoms to be, in general, slightly higher than the values found in this sub-study. It would be expected that case notes would include more information about a patient than an e-referral to specialists. Also, given that Hsiang et al. (2013) inspected the case notes manually, presumably the presence of symptoms in patients was assessed with very high accuracy.

With respect to combinations of symptoms presented by individual patients, the most commonly reported combination in this sub-study was change in bowel habit and rectal bleeding, which is consistent with data presented by Hsiang et al. (2013). Another common combination of symptoms in this sub-study - abdominal pain and weight loss - was also previously reported in another NZ study by Windner et al. (2018), who used patient reported symptoms in their analysis. The number of patients with CRC presenting with more than one symptom was much higher in Windner et al. (2018) than in this sub-study (72% vs 50%), which is a consequence of using patient-reported symptoms.

The main aim of this sub-study was to develop a statistical model for calculation of CRC risk in individual patients referred to specialists for further investigation of presence/absence of CRC, based on information included in e-referrals. The presented final model with shrunk coefficients (in order to make the model more realistic for predictions of CRC risk in future patients) was internally validated using cross-validation. Based on the predictions of CRC risk from the cross-validation procedure, I found that the 20% of colonoscopies were performed on patients with very low (less than 1.5%) CRC risk.

However, out of 660 patients with very low risk of CRC, four actually were diagnosed with CRC. Those four patients were all females diagnosed with localised cancers in stage I. Three of them were younger than 60 years, three had tumours in sigmoid colon and one in proximal colon, and the symptoms which they presented were nearly all negatively associated with CRC.

The final model included the following predictors associated with CRC: anaemia; palpable mass in abdomen; rectal bleeding; weight loss; patients' age at referral; and gender. The model also included the following predictors negatively associated with CRC: abdominal pain; family history of CRC; and IBD.

Anaemia (any type) was found to be an independent predictor of CRC, with adjusted OR=2.66. Unfortunately, it was not possible to assess the risk of CRC in patients with iron deficiency anaemia. A study by Panzuto et al. (2003) reported OR=10.4 for iron deficiency anaemia, which is much higher than the association found in this sub-study for anaemia of unspecified type. The diagnostic value of iron deficiency anaemia in our cohort may well be higher than the estimated value of unspecified anaemia, but the distinction was not reported systematically in the free-text notes and is therefore not possible to estimate from the data. **Palpable mass** in the abdomen or rectum had adjusted OR equal to 2.56, which is low when compared to the tabulations reported by Chohan et al. (2005), according to whom the risk of CRC in UK patients referred to colorectal clinic, who presented with an abdominal or rectal mass, versus those without a mass, was OR=15.4. The much lower OR in my study could be due to the exclusion of those who did not have colonoscopy (who were included in the UK study). As a result, almost all patients in my study had fairly high CRC risk, while the UK study included many patients with low CRC risk. Therefore the UK study has many true negatives, which leads to a high OR. Eight out of 50 patients with a palpable mass were diagnosed with CRC, which means this symptom has a high PPV. Despite the high PPV for palpable mass, as can be seen in Figure 5.10, the sensitivity is very low, which makes the symptom

useless as a single predictor (too many patients would be missed).

Other things being equal, the risk of CRC in patients with unspecified **rectal bleeding** was higher than in those without rectal bleeding, which is consistent with findings from previous studies (Jellema et al., 2010). I also investigated dark blood in stool as a separate symptom, and found a crude OR=2.14. However, the association between CRC and dark blood in stool was not statistically significant in the adjusted model as the prevalence of dark blood in our cohort is low (2%) and the study cohort was too small to show the effect of symptoms with such low prevalence. Association with **dark blood** in secondary care was reported earlier, however, with crude OR almost twice higher than in this sub-study (Bjerregaard et al., 2007). **Weight loss** in this sub-study was a statistically significant predictor of CRC only in terms of crude OR. When adjusted for other symptoms, age and gender did not reach statistical significance: however, based on the backward elimination procedure, weight loss was left in the model. I hypothesised that the lack of statistical significance could be due to confounding between weight loss and anaemia, but this was not the case, as there was no association between those two symptoms (OR=0.97). This result is similar to results by Adelstein et al. (2010) who reported a very similar crude OR as in my analysis, but weight loss was not included in their final model.

Family history, abdominal pain and diarrhoea were all associated with lower risk of CRC in this sub-study. The finding that **abdominal pain** was negatively associated with CRC is in general consistent with literature (Jellema et al., 2010). Surprisingly, however, Adelstein et al. (2010) found a positive association in those patients who had weekly abdominal pain for less than 12 months, and no associations for longer duration and lower frequency. In contrast, another study which analysed data from refereed population in Norway (Steine et al., 1994), reported no association between duration of abdominal pain and CRC. Information about duration and frequency of symptoms was unfortunately not included systematically in the free-text notes analysed in this sub-study. Similarly to what this sub-study found, a lower risk of CRC in patients with abdominal pain has been

reported in earlier studies (Steine et al., 1994; Selvachandran et al., 2002; Bjerregaard et al., 2007; Thompson et al., 2008). Because the comparison of the adjusted associations is not straightforward (different studies adjusted for different predictors), I compared crude ORs for abdominal pain. The magnitudes of the negative association in those studies (crude ORs: 0.39; 0.43; 0.73; and 0.59 respectively, based on values reported by Jellema et al. (2010)) were similar to what I found [crude OR=0.44 (95% CI; 0.27, 0.71)].

While there is generally a consensus about the negative association of abdominal pain and diarrhoea with CRC, there is a bit of controversy about family history of CRC. The lower risk of CRC in the referred population of patients with family history, found in this study, was also reported in an Australian referred population (Adelstein et al., 2010, 2011). However, Bjerregaard et al. (2007), based on analysis of data from Danish population that was comparable to this sub-study with respect to prevalence of family history and CRC, reported OR=2.4, showing a positive association, in contrast to the result from my analysis (OR=0.46). Due to the conflicting evidence, the NICE guidelines for the "two week referral" have not included family history as a high-risk criterion for further investigation of CRC (Thompson, 2002).

While such a negative association for those three symptoms may appear contra-intuitive, especially the association with family history of CRC, which is clearly positive in the general population, the phenomenon is, however, easy to explain. The negative association is based on data from the referred population. Due to the positive association in the general population, GPs presumably refer too many patients with family history, but without any symptoms which could suggest CRC. As a consequence, many of the referred patients with a family history will not have CRC, thus causing the negative association.

Age was the most influential risk factor for CRC. For example, the adjusted OR was 1.68 for an increase in age from 60 to 70 years. The OR was lower than the IRR in the

general NZ population in the year 2016, as reported by Ministry of Health NZ (2018a) (IRR=2.33 for between 60-64 and 70-74 age brackets) and for the year 2018, as found in this study (IRR=2.00 for Māori and IRR=2.25 for non-Māori for 70 years relative to 60 years, based on the values in Figure 3.25 in Chapter 3). The lower OR is due to young people in the referred population having a higher CRC risk than young people in the general population, as the young individuals in my study have been referred due to symptoms associated with CRC, and therefore they have much higher risk.

Due to the strong association between age and CRC risk, the model predicted CRC risk in the oldest patients is high even in the absence of symptoms. For example, a male 85 years old, without any of the symptoms included in the final model has a CRC risk of 12% which is similar to the CRC risk for ∼62 years old male presenting with rectal bleeding and weight loss (Figure 5.9).

Other things being equal, males had much higher risk than females (OR=1.6). Within the study cohort, fewer males than females had a colonoscopy (42% were males), although the majority of CRC patients were males (54%). Further, there were relatively many women who had a colonoscopy despite having low CRC risk: 33% of women with CRC risk lower than 2%, compared to 20% of males. This suggests that males underutilise colonoscopy resources. One factor that could explain the low number of colonoscopies in males is relatively few males being referred by GPs; secondly, after referral specialists appear not to consider male gender to be a strong risk factor.

With respect to the investigation of factors which influenced time from first referral to colonoscopy, patients with symptoms not associated with CRC: haemorrhoids, anal symptoms, abnormal liver functions and abdominal pain waited longer for colonoscopy, while presenting with occult blood, anaemia, rectal bleeding, and older age was found to be associated with shorter time to colonoscopy. This suggests that specialists give priority for earlier colonoscopy to patients who are at high risk of CRC, and that the

findings of this sub-study are broadly consistent with the expertise of doctors in Waikato Hospital in Hamilton. Median time to colonoscopy was longer for Māori but, in the Cox model, adjusted for age and symptoms, the association between ethnicity and time to colonoscopy was far from statistical significance. The crude association could be related to the younger age of Māori in our cohort or could be due to the different prevalence of other predictors in those two ethnicities.

5.5.2 Strengths

An important strength of the study is that all patients included in the study cohort underwent the same reference test (full colonoscopy) to diagnose CRC, an important factor to consider when evaluating the performance of a model for risk assessment (Bossuyt et al., 2003). Secondly, all patients who satisfied the inclusion criteria were included in the statistical model. Third, the extraction of symptoms using the automated procedure achieved high accuracy of extraction. I estimated that over 90% of patients had accurately extracted symptoms. Finally, the model for calculation of CRC risk for assisting secondary care doctors was based on the same information as what is available to physicians in e-referrals.

5.5.3 Limitations

The study has several limitations which have to be acknowledged.

Firstly, using the same reference test for the whole cohort, on one hand, is a strength, and prevented the results from being affected by verification (work-up) bias (O'Sullivan et al., 2018). On the other hand, the choice of full colonoscopy as a reference test for the whole cohort created a limitation (selection bias). Due to the exclusion of patients who did not have a full colonoscopy, the study cohort was not representative of the

target population with respect to age and ethnicity. Māori and younger patients were underrepresented.

Second, although use of the automated procedure to extract information from free-text notes limited observer bias, I would like to acknowledge some possible subjectivity, which was unavoidable as only one person (myself) took decisions about key words included in the code for symptom extraction.

Third, the referrals rarely specified that a patient did *not* have a specific symptom. Whether the lack of specification in an e-referral of a particular symptom was due to missing information (e.g., the doctor did not ask), or due to genuine lack of the symptom, is difficult to know. Considering the report by Selvachandran et al. (2002) that a patient questionnaire provided more information about patients' complaints than GPs' referral letters, comparison of my results to results from studies which used a different design, e.g. a questionnaire, to assess presented symptoms, should be done with caution. This is because the estimated coefficients in this sub-study might be underestimated compared to what they would have been if the symptoms had been elicited more accurately (whether due to GPs' elicitation of the symptoms or to the extraction of the symptoms from e-referrals).

The small number of patients included in the study cohort is another limitation of this sub-study. Many patients from the study population did not have any symptoms specified (55%), and so could not be included in the analysis. Due to a low number of patients diagnosed with CRC in the study cohort, the more complex prediction model - specifically the model with biologically plausible interactions (age/gender/anaemia) - was not supported. Also, the small number of CRC patients who satisfied the inclusion criteria could be a plausible reason why I could not show, in adjusted models, any effect of ethnicity on CRC risk and on the time to colonoscopy. Since differences between ethnic groups with respect to the quality of, and access to, medical care is a central topic in NZ

public health debates, and earlier studies have shown differences between ethnic groups in NZ with respect to diagnosis of CRC and outcomes in CRC patients (Hill et al., 2010; Sharples et al., 2018). It would be valuable to address this issue, which unfortunately was not possible in this sub-study.

Regression trees turned out not to be useful for identifying interaction terms for inclusion in the model, as none of the proposed interactions improved the predictive accuracy of the model. The most likely reason is that the sample size was too small to determine important interactions. Some interactions between symptoms are known to be biologically plausible. An example is the effect of anaemia on CRC in different age groups in women (Hamilton et al., 2005), but I was not able to verify this interaction term due to the small number of young women in the sample.

5.5.4 Data quality

The analysed data sets contained no inconsistencies with respect to gender, date of birth, or dates describing chronology of the following events: referrals; colonoscopies; and cancer diagnoses. However, the recorded number of colonoscopies performed on individual patients was sometimes very high (in the most extreme cases, up to twelve colonoscopies in a short period of time) raising the question what the reason was: was it the same colonoscopy erroneously recorded multiple times, or had the patient really had such a large number of completed colonoscopies? We did not have opportunity to discuss with specialists the reasons for multiple colonoscopies and to correct the data if necessary. However, in this study it was only important to know whether a patient had had a full colonoscopy, and the number did not affect the presence or absence of the reference test. A similar issue applied to the number of referrals per patient. Here, however, the reason was usually clearer, and it was easy to detect those multiple referrals which were sent or entered multiple times by an accident.

The free text comments were of poor quality due to the lack of a standardised form for specification of the symptoms. For example, in some comments it was not clear whether a symptom was indicated, or whether the specialist was asked to investigate the presence of the symptom (the free-text contained many question marks and it was not obvious whether the symptom were present, e.g., "pain in abdomen?"). Because the comments were assessed by only one investigator (MH), there is a risk of observer bias. The lack of a standardised form of specifying symptoms in e-referrals makes it difficult to use the data for scientific research (Koeling et al., 2011). There is probably no easy solution to this problem, which was investigated previously by Resnik et al. (2008). The authors argued that, although entering the information by clinicians (physicians) in a more structured way could be a better solution, in general, physicians need the flexibility which is offered by free text fields in clinical data bases, and it might be especially important when entering symptoms which have not been coded yet, or for the description of the severity and extent of symptoms.

Although the research carried out by Resnik et al. (2008) was based on research data bases, the same need applies to e-referrals, possibly even to a greater extent. This is because e-referrals, in contrast to research data bases, are made for the reason of medical care for patients, not for scientific research. It is therefore important for GPs to be able to express in their own words the health problems of patients for whom e-referrals are made, and to use their own judgement and expertise in explanation. A possible - and probably the best - solution would be to develop freely-available high-quality software for the extraction of information from free-text notes in e-referrals. Such software would need to be tailored to the specific medical terminology, in this case lower gastrointestinal disease.

Another issue with the free-text data was the lack of systematic information related to: duration and frequency of symptoms; degree of the symptoms; degree of weight loss; and specification of whether anaemia is due to iron deficiency. As shown in Adelstein et al.

(2010), such distinctions might give valuable information.

5.5.5 Relevance to clinical practice

Although I analysed data from a single hospital only, important data about CRC in a sample from our local population are presented which can have implications for clinical practice. Due to a screening program which will be implemented in Waikato region in 2021, some patients aged 60 to 74 years will be diagnosed via the screening pathway, but the findings will still have implications for all those patients who are outside of the screening age range, as well as for those who will not participate in the screening program. The New Zealand participation rate for Round 1 of the pilot screening program was 56.9%, and 51.6% in Round 2 (Ministry of Health NZ, 2016). It is expected, therefore, that nearly half of the eligible population will not be screened, and therefore, for those people, diagnosis of CRC will have to rely on symptoms and other predictors. Additionally, in our cohort, 17% of CRC patients were younger than 60 years, so they could be diagnosed only as symptomatic patients.

The statistical model for calculation of CRC risk developed in this study was, when the study was designed, meant for use in the secondary care settings to assist specialists in decision making for selection of patients for colonoscopy. After conducting the study, I found that the decision is made in two steps of the diagnostic pathway. In the first step, the specialists have to decide about accepting a patient for FSA, and in the second step, the actual decision about colonoscopy is made. The model is best suited for assistance in the first step as it can be assumed that specialists take the decision about FSA based on information included in e-referrals. However, for this purpose, the predicted risk may be overestimated as the model is based on patients who got colonoscopy, and in this population the risk of CRC is higher than in the whole referred population. Moreover, I did not include patients who did not have any symptoms specified in the e-referrals. This

choice was made because I assumed that the decision about selection of a patient for FSA is based on presented symptoms, and therefore the specialist most likely had access to information not included in the e-referral. Therefore, in order to obtain a representative sample, it would be necessary to get access to the communication between primary care doctors and the specialists.

Secondly, the model could be used for making decision about selection of patients for colonoscopy. However, in that step the use of the model is more limited, as the specialists use additional information that transpires during the FSA. This information was not available for this sub-study. Therefore, to make the model suitable for use in clinical practice while making decision about colonoscopy, the fitted model should ideally include information from clinical notes gathered during FSA.

I have to acknowledge that the saving of 20% of performed colonoscopies during the study period is not achievable in practice as many of the patients who got colonoscopy were assessed based on additional information not only on the e-referrals. When designing the study I did not realise that the decision for performing colonoscopy is taken based on combined information from e-referrals and from the information gathered in FSA.

Also, the evaluation of the performance of the model is valid under the assumption that doctors make decisions about colonoscopy specifically for investigation of CRC. In reality, some patients might have a colonoscopy to investigate presence of other disease, such as IBD alone, or in addition to suspected CRC. It is therefore possible that, among those 660 patients who, according to my model, did not need colonoscopy due to the low CRC risk, some patients actually should have a colonoscopy for other reasons, despite the low risk of CRC.

This section concludes the empirical part of the PhD project. The next chapter synthesises the findings and provides a discussion relevant to the whole project, as well as closing remarks.

Chapter 6

Overall discussion and conclusions

This chapter discusses the findings from the empirical part of this study (chapters 3, 4 and 5), analyses their significance, and proposes potential next steps suggested by this research. This chapter opens by recapitulating the gaps in knowledge at the onset of this research. I then synthesise the findings, and discuss their validity, in order to show how this PhD project contributes to the body of epidemiological knowledge. Subsequently, I discuss how the findings apply to clinical practice in NZ. Finally, I summarise the strengths and limitations of the methodology used in this study, and discuss directions for future research.

6.1 Introduction

At the beginning of this study, I hypothesised that it will be possible to provide valuable epidemiological evidence related to the incidence of CRC in NZ, as well as the risk of CRC in NZ patients, based on analysis of already existing data. I further suggested that such results could be useful in tackling the burden of CRC in NZ. Analysis of existing data to uncover new knowledge is, in the case of CRC epidemiology, important and relevant for the following two reasons: first, existing data contain evidence relevant to all NZ patients without the need to place additional stress on patients by collecting new data using patient questionnaires; second, the MoH in NZ holds good-quality population-based registry data sets that are available to researchers and can provide results of high scientific value.

Among the identified gaps in the NZ research related to CRC epidemiology, I chose three that can be addressed by the analysis of already existing population-based and administrative data: analysis of the trends in CRC incidence using age-period-cohort models, including a separate analysis for Māori and non-Māori populations; investigation of the association between diabetes and CRC in NZ patients; and, finally, as I was given the opportunity to analyse data from patients referred to secondary care (Waikato DHB)

as a part of the HRC study, I developed a model for CRC risk in individual patients who were referred for investigation of presence/absence of CRC.

In this thesis, I have addressed all three objectives. The results of the statistical analyses that I carried out support my hypothesis, and the analyses uncovered interesting and important information. The next section provides a synthesis and discussion of the most important results.

6.2 Synthesis and discussion of main findings

I consider the identification of the complex cohort effects in the CRC incidence data as the most important finding from this study. The cohort effects can explain the whole dynamic in CRC incidence in the NZ population between 1994 and 2018. The interpretation of the results from the APC model is subject to the assumption of zero period slope, which, in the case of CRC in NZ, is a biologically plausible assumption (as explained in section 3.1.2.3). Further, following the argument by Dobson et al. (2020), allocating the drift to the period effect could mask a cohort effect, which is important information for clinical practise and health promotion.

The APC analysis revealed that generations of non-Māori born around 1940–1955 (Figure 3.8) experienced a steep decrease in incidence rate ratios with increasing year of birth. Further, the incidence rate ratios in generations born in the 1970s and 1980s increased steeply with increasing year of birth, regardless of gender, and to a similar extent in Māori and non-Māori populations. The increase in IRRs applies to all three anatomical sub-sites: proximal colon; distal colon; and the rectum with one optimistic feature; the trend in distal cancers showed a sharp downwards turn in generations born in the 1980s, after the strong increase in IRRs in those born in the 1970s. Due to the wide 95% CIs, however, the trend has to be taken with caution. Despite the uncertainty of the observed

trend, it is interesting what could cause the possible sudden change in the direction of the cohort effect in generations born in the 1980s solely for distal tumours.

Generations born from around 1980, following the definition of cohort effect, were exposed/not exposed in early life or puberty to some specific factors associated with (or protective against) risk of developing distal tumours, e.g., lifestyle or nutritional factors. Considering that one of the lifestyle factors associated with CRC, physical activity level, has been shown to be associated with both distal and proximal colon cancers (Boyle et al., 2012), and that diabetes, another possible risk factor, was found in this study equally to affect incidence of proximal and distal tumours in the NZ population (presented in Section 4.3.5, Q5), dietary factors may be a possible reason. An example is high red meat consumption, reported to be associated with increased incidence of distal tumours (Norat et al., 2005). Because meat intake in NZ decreased between 1975 and 1990 (Laugesen, 2000), reduced meat intake is an example of a factor that can be directly or indirectly related to the decreasing trend in IRRs for distal tumours.

Despite the small optimism related to this reversed trend in incidence of distal cancers, the pattern revealed by the APC analysis is alarming. This is because the very steep recent increase in incidence rates in young New Zealanders can, nearly entirely, be explained by the cohort effect, which implies that the increased risk will follow the affected generations throughout their lives. As a result, in the near future, when the young generation affected by high CRC incidence rates replaces the older generations with low rates, a wave of CRC diagnosis can be expected in NZ; this finding is a confirmation of what was announced by Christopher Jackson (Cancer Society NZ, 2019), based on the results from Araghi et al. (2019). However, based on the supplementary tables from Araghi et al. (2019) (obtained from the authors), the IRRs differed substantially from those estimated in this study. For example, the IRR for colon cancer in the 1980 birth cohort, compared to the 1945 cohort, was 0.88 in Araghi et al. (2019), compared to 1.85 in this study.

I would argue that the results reported in my study are more reliable, for the following reasons: the estimates from this study are based on newer data (the Araghi et al. study analysed data up to 2014); further, Araghi et al. used the default option of modelling timescales using 5 knots, which, for the analysis of longer incidence periods, is not optimal. When I fitted the model using the default set of knots, the model provided very poor fit to the NZ incidence data, and Araghi et al. (2019) did not provide any evidence for the fit of their model to the NZ data. In this study, I carefully designed the APC model by selection of the knots for the splines based on backward elimination, rather than using the arbitrary default set of knots.

Modelling of the time-scales (age, period and cohort) using splines made it possible to capture the complex effects of time-scales, while keeping the fitted model simple enough to allow analysis of data from the Māori population, a stratum of the NZ population with a small number of CRC patients, with satisfactory width of 95% CIs. It was the first study to attempt an APC analysis of CRC incidence in Māori and non-Māori populations. The study provided a description of trends in CRC incidence in the period 1994–2018, and it described the contribution of age, birth cohort and year of diagnosis to the incidence of CRC. The APC analysis stratified by ethnicity uncovered several important and interesting issues which are discussed in the next paragraphs.

Among patients diagnosed with CRC, Māori ethnicity was undercounted in the NZCR up to the year 2006. This undercount must therefore be corrected when analysing trends in CRC incidence by ethnicity, if using the count of population obtained from the census as the denominator. For the models stratified by ethnicity, I conducted two analyses: one analysis which used data without any correction for the undercount (Section 3.3.3.4); and a second analysis which included the corrected counts of Māori and non-Māori CRC cases (Section 3.3.5.1) based on correction factors provided in the literature (Shaw et al., 2009; Boyd et al., 2016). For both data sets (with uncorrected and corrected counts), there was no statistically significant disagreement between observed and predicted IRs.

The correction factors for the count of Māori patients in the NZCR for years before 2006 are extremely high and, therefore, as expected, the cohort effects and the net drifts from both analyses were different, and it is not obvious which of the two models is more plausible. However, when inspecting figures with ASRs based on data with uncorrected counts (Figures 3.2 and 3.3), and on data with corrected counts (Figure 3.21), the rates based on uncorrected counts show a flat trend between years 1994–2018 for Māori, which seems to be more plausible. The rates based on corrected counts (Figure 3.21) are high in years with high correction factors, which suggests that using correction factors leads to over-correction of the count for Māori patients diagnosed with CRC. Because the undercount of Māori patients with CRC and the true ASRs in Māori are two unrelated processes, no correlation between a corrected ASR and correction factors would be expected, if the correction factors were correct. I acknowledge that this is not a very strong argument, since both processes are confounded with time, and therefore some correlation is not surprising. However, there is no a priori reason to expect a positive correlation, or even to expect that the two processes would jump in the same years.

Further, the decrease in the size of the correction factor from year 2004 to year 2006 is dramatic, from 1.31 to 1.01 (for years 1997–2000 the correction factor is as high as 1.48). The question of whether such a sudden improvement of severe undercounting - going from 1.31 to 1.01 - is ever possible within such a short period, is important to ask here. Considering that the records of ethnicity in NZCR since 1999 are based on NHI recordings that can be several years old (Shaw et al., 2009), it does not seem plausible that the undercounting could change so abruptly from year to year. Rather, a smooth change would be expected as the old NHI records are gradually replaced by new records.

I would like to acknowledge that the work done by Shaw et al. (2009) and Boyd et al. (2016) is impressive and very important. However, how reliable the estimated correction factors are is impossible to say, as, according to Boyd et al. (2017), there is no very good method to estimate the accuracy of the probabilistic linkage which both authors

used in their work to calculate the correction factors. Further, the use of highly probable links, which Shaw et al. (2009) used in their work, resulted in only 70% to 76% of the NZCR records being linked to census records in the period relevant to this study. The undercount of Māori ethnicity in the remaining records is not known. It is therefore possible that the correction factors are influenced by selection bias.

With respect to ASRs, regardless of the analysis I consider (with uncorrected or corrected counts), in the period 1994–2018 there is no increase in ASRs in Māori. According to my analysis, in the 25-year study period, based on the data without correction for the undercount, the ASRs in Māori did not show any consistent trend. By contrast, using data corrected for the undercount, the ASRs decreased nearly as fast as in non-Māori. As can be noted here, the trend in ASRs when using data corrected for the undercount, gives a picture opposite to what is reported in NZ publications, namely that the rates in Māori are growing faster than in non-Māori.

Due to such problems, it is not possible for me to say which description of trends for Māori populations is more likely to be closer to the true trends, and which cohort effect is more plausible; either of the two would be only speculative. To answer the question, it is important to undertake a collaborative project which would involve scientists with a broad knowledge of the history of the NZ population born from 1900 or even earlier, including knowledge of the most accurate methods for correcting the undercount of Māori ethnicity in the NZCR, as well as epidemiologists or biostatisticians with the skills to analyse data using advanced statistical models. It is extremely important for NZ to carry out such a project, as it could answer very important questions related, not only to CRC, but generally to many health outcomes in Māori populations, as other forms of health research are likely to have similar issues with recording of ethnicity.

The description of the trends in IRs in Māori, one of the main objectives of this study, thus cannot be fully resolved in this project. I provide, however, the results given by two

models, which shows the magnitude of the problem in providing reliable estimates of the trends in Māori populations. The trends from 2006 onwards may be more reliable and, when more data becomes available, conducting an APC analysis restricted to data from after 2005 should provide more reliable estimates. However, unless the issue with historical data is resolved, many interesting research question may remain unanswered.

Another main finding came from the investigation of the association between diabetes and CRC (sub-study 2). I found a statistically significant association between diabetes and CRC in the NZ population, although weaker than the association reported in most other countries. Although methodological differences between studies could account at least partially for the disparity, the difference between NZ and other countries could be real, especially considering that, even within one country, as shown in this study for NZ data, the diabetes-CRC association appears to differ between ethnicities. While, in the Māori population, diabetes was not associated with CRC, in non-Māori an association between diabetes and CRC was identified, but, interestingly, only in individuals younger than 75 years. The interaction between age and diabetes is an important finding in its own right, and also helped me to discover one possible reason for the differences between studies with respect to the strength of the association, namely the difference in age distribution between studies. This finding was possible due to modelling of the age effect on IRs using splines, which is not a standard investigation in studies on associations diabetes-CRC.

Although the results of the analysis indicates differences between Māori and non-Māori with respect to the association between diabetes and CRC, I can not draw any conclusions (based on the analysed data) about possible reasons for the difference. The investigation of such reasons requires deep knowledge of health determinants in the NZ population and access to relevant confounders (Cormack et al., 2019).

The last result which I consider as important is the developed model for the CRC risk in

individual patients referred to secondary care for further investigation of presence/absence of CRC. The explanation of how the model could be incorporated in clinical practice is given in Section 6.6.

In addition to the main findings presented in this section, the study has several additional findings which are briefly discussed in the next section.

6.3 Summary of additional results

This study addressed an issue raised in the report to the HRC (Secker et al., 2015), namely, the need for investigation of how the difference in age at diagnosis in Māori and non-Māori populations can be explained. According to the results of the APC analysis, using uncorrected counts of CRC cases by ethnicity, the difference in median age at diagnosis, which, for the population studied in sub-study 1, is around 7.5 years, can be explained mainly by the fact that the Māori population is younger, partially also by the cohort effect (two years) and, only for a very small part (48 days), by age effect. However, based on an analysis of data with corrected counts, approximately two years of the difference in median age at diagnosis is explained by age effect, and very little by cohort effect. The problem was not solved previously as, in order to draw longitudinal age curves, the age effect has to be adjusted for cohort and period (Bell and Jones, 2013) which can be achieved by carrying out APC analysis.

The APC analysis confirmed the relatively high IRs in proximal tumours in older females reported previously in NZ literature (Shah et al., 2012). This study adds to the already known phenomenon, that it is entirely due to the age effect and therefore can be expected to persist in the future. Additionally, the left- to right-sided shift reported by (Shah et al., 2012) was also confirmed by the results of the APC analysis. This result shows that the shift is explained entirely by cohort effects related to generations of females

born between 1904 and 1939 (Figure 3.17). For females born between the years 1940 and 1980, the cohort effects for proximal and distal cancers are very similar and, therefore, in the near future, when the CRC incidence will be dominated by those generations, the left- to right-colon shift is expected to disappear. However, according to the model, the downwards turn in the cohort effect for distal tumours in generations born after 1980, is expected to cause the shift to return in the distant future, and to affect males and females.

According to the results of the adjusted Cox model for time from referral to colonoscopy, the waiting time did not differ between Māori and non-Māori populations. The median time was 20 days longer in Māori than in non-Māori, but this difference was most likely due to the lower median age in Māori patients and the negative association between age and time to colonoscopy.

Before I discuss contribution of the study to the body of knowledge and applicability of the results to clinical practice, in the next section I discuss the validity of the results.

6.4 Validity of the study results

Sub-studies 1 and 2 were observational studies which analysed administrative and population-based registry data, which protects the external validity of this study, that is the applicability of the study results to the target population, subject to the internal validity (Szklo, 1998; Grimes and Schulz, 2002). To ensure that the internal validity is satisfactory, I addressed several biases that could have distorted the results. In addition, the quality of analysed data have an impact on the validity of results. Below, I firstly discuss possible problems with validity due to the data, and, afterwards, I state additional biases and problems with applicability of the results.

In general, the accuracy of the NZCR is good and, as reported by Cunningham et al.

(2008), appears to be similar to that found in comparable audits of cancer registries. The stage at diagnosis has the lowest accuracy among included variables, similar to elsewhere (Gatta et al., 2003). This study's results were not affected by this inaccuracy, as I used stage information only for descriptive statistics. Another problem with the data from NZCR is the undercount of Māori ethnicity in registrations before the year 2006, affecting the data analysed in sub-study 1 between years 1994 and 2005 (Boyd et al., 2016). The issues with quality of the ethnicity information in NZCR caused the results of the analyses by ethnicity in sub-study 1 to be affected by misclassification bias. As shown in the analysis using corrected counts, the misclassification bias has a strong influence on the net drift and on the predicted IRs before 2006 in Māori populations. The bias has also some influence on the cohort deviations, period deviations and longitudinal age curve in Māori populations. The issue is discussed in length in section 6.2.

To avoid the problem, a better choice for the analysis would be the use of a probabilistic linkage of NZCR and individual census records as e.g., in Shah et al. (2012), which allows the use of census ethnicity for both numerator and denominator. I did not attempt to implement this approach as, when I discovered the problem, it was too late to incorporate into the method of the PhD project, as my ethics approval did not include access to individual census records. However, when reflecting on my approach, I found that the two analyses which I conducted and presented gave deep insight into the issue of how problematic the undercount of Māori ethnicity in NZCR is for the analyses of long-term trends in CRC incidence separately for Māori and non-Māori populations.

The other major data set analysed in this study is VDR. The main problem that could affect the results is selection bias present in the VDR, as only a portion of patients with diabetes are registered (shown by the sensitivity of 87% (Ministry of Health NZ, 2018c)). It would be expected, however, that, among NZ patients with diabetes, those diagnosed with CRC are most likely to be registered in VDR, as they have many health care episodes, which MoH uses to populate the VDR. This caused a misclassification

bias in sub-study 2 which I could not address and, therefore, under the assumption of non-random lack of registration, the estimated associations between diabetes and CRC (overall and in population strata) could be slightly overestimated.

The additional biases addressed in sub-study 2 include: **detection bias**, which was addressed by the inclusion of the duration of diabetes and through the analysis of incident and prevalent diabetes; and **immortal time bias**, which was addressed by treating a patient as an insulin-user from the date of the second redemption of insulin.

The application of the results from sub-study 3 to the target population (patients referred to Waikato DHB for investigation of CRC) has to be taken with caution, because the analysis was based on patients who had a colonoscopy, which is not a representative sample of all referred patients. Further, the results from sub-study 3 are not applicable to the NZ population as a whole. Although the public health system is the same in the whole country and the same referral guidelines apply, there might be differences in the way how GPs and specialists approach their patients in different NZ regions (Bagshaw and Ding, 2019). Also, as shown in NZ literature, access to primary health care differs by regions Cormack et al. (2005), which might be a reason that patients do not present to GPs with severe symptoms, but rather go to the ED, omitting the primary care part of the pathway.

6.5 Contribution to the body of knowledge

This research contributes to the knowledge of epidemiology of CRC, primarily in NZ but also worldwide, by providing:

1. An analysis of the trends in CRC incidence rates in the NZ population for the period 1994–2018, based on age-period-cohort analysis. The results include the description of the contribution of the age, period and cohort effects to CRC incidence for the

whole population and separately for Māori and non-Māori. The results are also provided for CRC overall, for proximal colon, distal colon and the rectum, stratified by gender.

2. The predicted incidence rates in the NZ population for the year 2018 based on age-period-cohort models for each of the three anatomical sub-sites, stratified by gender, as well as predicted incidence rates for Māori and non-Māori populations. These predictions are based on population data; hence they are suitable for use in NZ clinical practice by doctors, as well as by policy-makers for updating polices related to CRC.

3. Support for the view that early- and late-onset of CRC are the same disease, by demonstrating that increased/decreased incidence in early life (before age of 30 years) follows each birth cohort throughout the whole life.

4. Evidence that the difference in median age at CRC diagnosis between Māori and non-Māori can mostly be explained by the population age structure and only for a small part by cohort or age effects.

5. An estimate of the strength of the association between diabetes and CRC in the NZ population.

6. Evidence for different associations between diabetes and CRC in Māori and non-Māori populations.

7. The finding that the association between diabetes and CRC in the NZ population applies only to patients younger than 75 years.

8. Evidence that, in Waikato DHB, males underused colonoscopy despite their higher CRC risk compared to females.

6.6 Applicability of the results to clinical practice

The practical application of the results to clinical practice is not within the scope of this study. However, as I aimed at providing results that, in addition to advancing epidemiological knowledge, could also be applicable to decision-making in clinical practice, below I discuss the applicability of results to the three application areas: primary care practice; secondary care practice; and policy-making.

6.6.1 Primary care

As explained by Sackett et al. (1985), results from epidemiological studies based on a group of patients can provide a valid basis for making inferences about an individual patient. Such inferences mirror what physicians do when they use their medical experience, based on patients previously seen in their medical practice, when making diagnoses in future patients. The use of predicted incidence rates based on data from the general population is especially important in countries like NZ, where a database with primary care information does not exist.

Therefore, the predicted IRs from the fitted APC models could potentially be useful to NZ GPs. However, they have to be presented in an appropriate way, as the standard presentation of the results of APC analysis is difficult to interpret (Smith et al., 2016). Especially difficult is the comparison of cohort effects in different population groups, as the interpretation of the results depends on parametrisation.

In this study, therefore, I propose a comparison of the predicted cross-sectional (for year 2018) IRs based on the fitted APC models, rather than a comparison of the cohort effects alone (as the predicted IRs do not depend on the choice of parametrisation). Using this approach, I provide the following comparisons: for males and females for the three anatomical sub-sites (Figure 3.18); and for Māori vs non-Māori (the graph for year 2018

in Figure 3.19). This presentation provides a suitable form for dissemination of the results of APC analysis to health professionals: in particular, it allows for an easy comparison of the IRs based on the fitted APC model (and therefore using many years of data for estimation of the IRs for a single year) in a simple graph over the whole range of ages, as well as a direct comparison between population strata.

The proposed diagrams could be used by GPs to assess the baseline risk of an individual patient based on the patient's gender and age. The diagrams, prepared separately for three anatomical sub-sites, could be helpful in choosing the type of investigation (colonoscopy vs sigmoidoscopy or digital examination) to manage existing colonoscopy resources better. It should be stressed that GPs will always have additional information about individual patients (e.g., symptoms and medical history), and this study does not address how such information can be combined with the baseline IR (for an overview see, e.g., Sox et al. (1988)).

The results of the APC analysis showed that, in 2018, IRs in Māori between 50 and 70 years old are already higher than in non-Māori of the same age (this result does not depend on whether data with or without correction for undercount were used). It is important for doctors to be aware of the relatively high incidence rates in Māori 50-70 years old. This is because, treating Māori patients as having lower CRC incidence could potentially harm Māori. If doctors apply knowledge based on older studies, if there are two otherwise similar patients who differ only with respect to ethnicity, applying the rule of prioritising the investigation of patients with higher risk - the non-Māori would have priority. Based on results of the APC analysis, this is an incorrect approach, as Māori between around 50 and 70 years old have IRs that are already higher than non-Māori, and only the oldest Māori generations have lower IRs. In that way, the finding from the APC analysis could contribute to a reduction of inequalities in health outcomes between Māori and non-Māori.

Based on the estimated strength of the association between diabetes and CRC, it is important to inform primary (and also secondary) care health professionals that the incidence of CRC in NZ patients with diabetes is only slightly higher than in non-diabetic patients. This information would enable doctors to apply evidence based on NZ data when dealing with patients, instead of relying on evidence gathered from different countries. To assist GPs during the assessment of a patient's need for further investigation of the presence/absence of CRC, the 13% increase in incidence in patients with diabetes may not be equally relevant to all patients. Diabetes status will be a less relevant predictor in patients who present with other risk factors (symptoms) that are correlated with diabetes status, and which are already taken into account by GPs during the assessment of CRC risk. An example of such a risk factor is constipation, a symptom that can be caused by both diabetes and CRC. In that case, the added value of diabetes as a predictor of CRC would be even lower than what the IRR estimated in this study suggests. Also, GPs should not consider diabetes as a risk factor for CRC in patients older than 75 years.

The lack of association between diabetes and CRC in Māori could have implications for choice of patients for further investigation; unlike in the case of non-Māori in whom diabetes status could be consider as a factor increasing the need for investigation in borderline cases, diabetes status in Māori patients, based on sub-study 2, does not seem to be relevant.

The increased incidence of CRC in insulin users, which is probably a proxy for the extent of diabetes and metabolic disease, could be considered as an additional factor for assisting clinicians in the decision-making process; for otherwise similar patients presenting with similar symptoms associated with CRC, an insulin user should have a priority for further investigation. The results, which indicate the increased incidence in insulin users, do not allow any inference to be drawn with respect to exogenous insulin being a causal factor for CRC development, and cannot be used to inform choice for anti-diabetic therapy.

Finally, this study found that males undergo fewer colonoscopies than females, despite the majority of CRC patients being males. Further, the detection bias in patients with diabetes was higher in males than in females. Because both findings suggest that males underutilise health services, doctors could encourage males to discuss any symptoms which might be related to gastrointestinal disease. Addressing this disparity between males and females in clinical practice may be one way to improve earlier diagnosis of CRC in the NZ population.

6.6.2 Secondary care

The main result of sub-study 3 is a prediction model for the estimation of the CRC risk in the referred population. ???However, the fitted model can not be used directly by gastroenterologists as a supportive tool for the selection of patients for colonoscopy because the model does not use data gathered during FSA, as data from clinical notes were not available to me. Also, the analysis did not include patients who had colonoscopy but did not have specified symptoms in e-referrals due to my assumptions that gastroenterologists had additional information about those patients e.g., from communication with GPs outside of the e-referral system. Having access to the communication between gastroenterologists and GPs would make the sample more representative.

The estimated CRC risk could ???then be used by gastroenterologists in two steps of the diagnostic pathway. First, it could be used while making a decision about the acceptance of an e-referral for FSA. Second, for those patients who were accepted for FSA, the calculated CRC risk could be used, in combination with other information which emerges during the FSA, for the selection of patients for colonoscopy. In both steps, the calculated CRC risk can be used by specialists as additional supporting information. Unfortunately, it is not possible to estimate how many unnecessary colonoscopies performed on patients with low risk can be saved, as, in practice, specialists who make the decision to refer for

a colonoscopy also use additional information that emerges during FSA, and the fitted model does not take that information into account.

In addition, young age was associated with longer waiting time to colonoscopy, which is of concern, as some studies (Chou et al., 2011; Mauri et al., 2019) have found that younger patients tend to present with faster-growing tumours. It is therefore possible that gastroenterologist should be encouraged to give, other things being equal, higher priority to young patients for timely investigation.

The low prevalence of dark blood in the referred population in our sample suggests that the dark blood is underreported by patients in primary care, probably because it is not easy to notice dark blood in the stool. If that is the case and dark blood is actually a good predictor of CRC, the advice included in NZ guidelines regarding discouraging use of FOBT test in symptomatic patients may be problematic, as it would be helpful for a gastroenterologist to know if a patient has occult blood, and performing the FOBT test would be the most reliable way to elucidate the information for a patient referred by GPs.

6.6.3 Policy-making

With respect to population-based screening for CRC, the results seem to be very relevant for application in future policy. As shown in Figure 4.13, if screening of patients with diabetes were considered at the time when they are at the same CRC risk as the general population at age 60 years, the screening age for those with diabetes would 57.5 years. Based on the predicted IRs from the fitted APC model for 2018 (Figure 3.19), the age 57.5 years would also be proposed for initiating of screening in Māori, as in the brackets of screening age, in 2018 Māori already had higher IRs than non-Māori. However, in 2019, Canterbury DHB (2019) wrote that "Bowel cancer is one of the few cancers for which Māori show lower registration and death rates than non-Māori". It is therefore

very important to analyse CRC incidence data carefully using appropriate models and to estimate IRs for relevant age groups, as such analysis gives more adequate and detailed information about IRs in the NZ population. The assessment of whether such changes to the screening policy are politically and practically feasible is beyond the scope of this study.

In addition to guiding screening policy, the knowledge that Maori do not have lower CRC incidence than non-Maori (except possibly for pre-WWII birth cohorts) but actually have higher IR than non-Maori (up to the age of about 70 years) is also important for information campaigns. The belief that Māori have lower CRC risk than non-Māori may lead to Māori being insufficiently attentive to symptoms related to CRC.

It is important for policy-makers to know that it is a cohort effect that explains the increasing incidence in young New Zealanders. The estimated cohort effect informs policy-makers about the expected wave of CRC diagnoses in the near future, as well as when the wave can be expected. This makes it possible to put in place policies that will allow the health system to cope with the coming increase in the number of CRC diagnoses, e.g., by ensuring more resources for gastroenterology services as, in NZ already in recent years, the capacity has been shown to be insufficient for current needs (Stamm et al., 2020). There is a need for prioritisation of resources to increase surveillance for CRC in those born in the 1970s and 80s, to allow easier access to colonoscopy which could lead to removal of polyps and diagnosis of CRC in early stages, when long-time survival is more likely. This is especially important for patients belonging to those generations as they are still young, and will therefore benefit the most from early diagnosis.

6.7 Dissemination of the results

Parts of the study results have been presented at conferences and in the report to the Health Research Council (as explained in chapters 3, 4 and 5). I am also planning to publish the results of the empirical parts of the study. Although during the Māori consultation (Appendix E) it was suggested to include Dr Nina Scott as an additional supervisor for the study, due to Dr Scott's high workload and later disruption related to the pandemic, this has not happened. However, as the feedback from the Māori Committee is very important, especially before publishing the results I contacted Dr Nina Scott after my thesis was approved for examination. During the meeting, it was agreed that the most relevant results that could help in addressing ethnic equity with respect to CRC outcomes in Māori, came from the age-period-cohort analysis (Chapter 3). Because the results from the age-period-cohort analysis can describe trends of CRC incidence very accurately, they can be used, e.g., during the discussion of the rational for early screening for CRC of Māori population. Appendix F presents a comparison of the age-specific rates from paper by McLeod et al. (2021) to the predicted rates by the APC model. As acknowledged by the authors, due to the strong fluctuations in age-specific rates, it was not possible to see how incidence trends differ between age groups for Māori and non-Māori. The presented figure shows how an APC model can provide a clear picture and thus help in the discussion.

6.8 Strengths

The main strength of this study was the use of population-based registry and administrative data sets, which possess the following advantages: reduced selection bias; large sample size; reduced burden on patients; reduced cost associated with data collection; data that have already been subject to data quality checks; and the possibility to carry

out long-term population-based studies in a short period of time.

Another strength was the choice of statistical models for the analysis in sub-study 1 and 2. In sub-study 1 used APC analysis, because the APC model is a simple model that, importantly, provided a good fit the the analysed data. Unlike joint-point regression, APC analysis allows information from all ages and all cohorts to be used in a single model. Use of generalised additive models in sub-study 2 made it possible to use multiple time scales (age, calendar time and duration of diabetes) without choosing a primary time scale, which could be a difficult choice to make when using survival analysis.

6.9 Limitations

Analysis of administrative data has, in addition to the strengths, some disadvantages which have to be taken into account. The main issue is that the administrative data have not been collected for scientific purposes according to a research protocol and therefore, baseline covariates (such as body mass index, diet, physical activity level which might be important in scientific studies involving non-communicable disease) are not recorded systematically. Also, there are changes to data collection methodologies over time (for example to the algorithm used in VDR and to the recording of ethnicity in NZCR) which can bias analysis involving time scales.

In this study, prioritised ethnicity was used throughout. This contradicts the recommendations from Statistics NZ (2005) and Didham and Callister (2012), which recommend total ethnic counts for use in epidemiological studies. I chose prioritised ethnicity because the Poisson regression used in the analysis requires the number of cancer cases per Lexis cell to be an integer. Further, double-counting individuals who identify as both Māori and non-Māori would give biased estimates, as the incidence rates for population strata with many double-ethnicity individuals would be inflated.

Further, the study did not provide an unambiguous estimate for the cohort effect in the Māori population which was one of the objectives of sub-study 1. This was caused by the problem with undercounting of Māori patients in the NZCR prior to 2006 when used in combination with census data as denominator. The results based on the use of published correction factors in Shaw et al. (2009); Boyd et al. (2016) does not seem to be plausible and the time-line for this project did not allow me to explore the topic further.

I would also like to acknowledge that the denominator used in sub-studies 1 and 2 is biased, as the denominator should include only the population at risk of CRC. However, the entire population count is used by all researchers for similar analyses, most likely because it is difficult to remove those who are already diagnosed with CRC from the population count tables. The bias is minimal due to the rarity of CRC.

Finally, although sub-study 3 provided a model for CRC risk in the referred population that has a potential to be used in clinical practice this study did not attempt to address the implementation of the model in clinical practice. To investigate how the model could be used by specialists requires a combination of qualitative and quantitative research. Proposing the model to doctors for their feedback, as well as validation in patients from other hospitals, is necessary in order to proceed further in applying the fitted model in clinical practice. This was not possible for this PhD study. However, the developed model is a good starting point as it has a potential to be helpful in clinical practice, assuming the further work will be carried out.

6.10 Directions for future research

The analysis carried out in this study filled a number of gaps in prior research; however, several new objectives emerged that could be addressed in future studies.

Firstly, repeating the APC analysis regularly every few years, when more recent data

become available, will provide an opportunity to monitor the future situation for individuals born in the 1970s and onwards with increased accuracy, as well as to assess incidence rates in the NZ population for those born after the 1980s, which this study did not include.

The results of the APC analysis are well suited to generate hypotheses about which risk factors could be responsible for the changes in CRC incidence between generations. Future studies could identify exposures that caused the strong decrease in CRC incidence specifically in successive generations of non-Māori born from around 1940 to 1960 and, secondly, could identify the subsequent exposures which affected generations born since 1970 regardless of gender and ethnicity. If the exposures could be identified, and they turn out to be related to lifestyle or pollution, health promotions and policies could help address the alarming trends in CRC incidence in future generations.

The research presented in sub-study 2 (association between diabetes and CRC) exemplified why modelling age on a continuous scale is important in studies on CRC associations. Modelling age using splines led to the finding that the association between diabetes and CRC relates only to patients younger than 75 years. It is therefore important that future studies in this area of research model age on a continuous scale. The example from my study could provide a teaching exemplar for young scientists to learn why analyses which are carried out using categorised age (especially when very crude categorisation is used, such as 10-year age bands, or dichotomised age) can lead to spurious results that lose important information about the age effect.

There is additional research question that emerged from the identified interaction between age and diabetes in sub-study 2. Future studies could investigate whether the diabetes-CRC association truly relates to those younger than 75 years, or whether the association affects only generations born after circa 1940. In other words, it would be important to investigate if this result is a cohort effect or an age effect. This study was unable to

answer the question, as the sub-study only analysed VDR data for those registered from 2014 to 2018; in such a short period, age and cohort are strongly confounded and it is impossible to distinguish the effects. As explained by Dobson et al. (2020), identification of a cohort effect in longitudinal data is important, as the knowledge would be useful for health promotions and for forecasting.

The results from sub-study 2 also generate a scope for the three following directions for future research: firstly, an investigation of factors that influence the association between diabetes and CRC differently in Māori and in non-Māori; secondly, an investigation of the factors responsible for the higher CRC incidence in insulin users by conducting analysis including duration of insulin use, the dose level, and relevant confounders; and thirdly, in the light of the findings from sub-study 1 with respect to cohort effect being responsible for trends in CRC incidence in the NZ population, it is possible that the interaction between calendar year and diabetes can be better explained by the effect of birth cohort and diabetes. In order to investigate a possible diabetes-cohort interaction a study similar to this sub-study but with much longer duration would have to be conducted.

Another opportunity for future studies would be to extend sub-study 3 to other hospitals in addition to the Waikato Hospital. The algorithm for extraction of information from free-text notes can be made accessible to other researchers on request. There are two reasons for the additional analyses: a larger sample could give a more reliable model for the prediction of CRC risk; and an assessment could be made whether one model would suit the whole country or whether different models should be used in different hospitals. However, before commencing such a study, a qualitative research project would be desirable, to interview gastroenterologists to investigate whether such a model would be suitable to assist their practice.

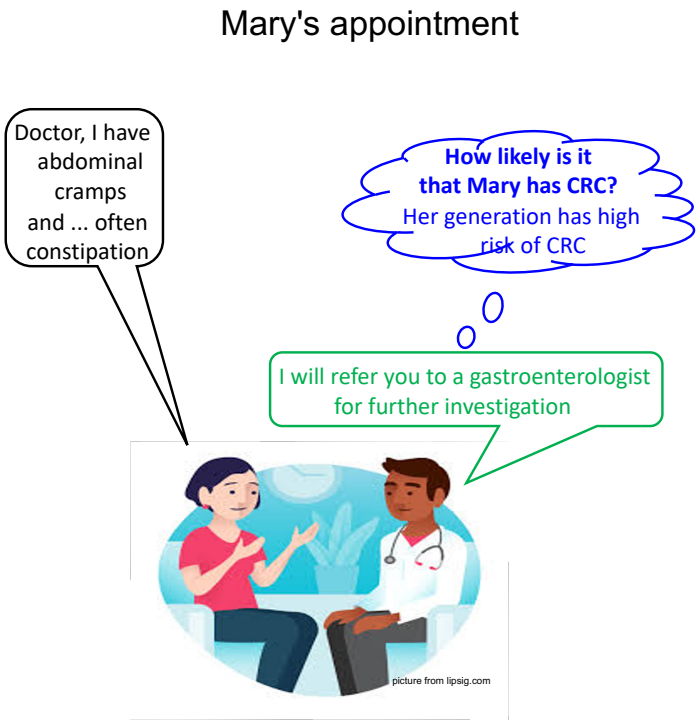
Finally, the APC analysis identified the need for conducting methodological research on the selection of knots for splines. As shown in this study, the choice of knots can

have a substantial impact on the model fit, and on the predictions. Implementation of appropriate methods for the selection of the number and location of the knots in software for fitting APC models might help to increase the uptake of APC models for analysis of longitudinal data.

6.11 Final conclussions

This study provides new insight into the CRC epidemiology in New Zealand - insight which has the potential to help patients, doctors and policy makers to make better informed decisions. In summary: generations of Māori born after circa 1946 no longer have a lower incidence of CRC than non-Māori from the same birth cohorts; patients with diabetes *do* have moderately increased incidence of CRC, but only non-Māori younger than 75 years; the increased incidence of CRC in generations born since the 1970s is unlikely to be limited to early-onset CRC, but will rather follow those generations throughout their lifespans; and, in secondary care settings in the Waikato DHB, males underuse colonoscopy resources despite having a higher CRC risk than females, while family history of CRC was found to be negatively associated with CRC, as patients with family history are more likely to be referred, despite low CRC risk. The findings may also help to formulate hypotheses that can identify environmental and lifestyle related risk factors which, one day, may be targeted to prevent unnecessary suffering from CRC.

These findings answer the question I asked at the beginning of the study: bringing back here our fictional patient Mary, I can confirm that doctors will be able to benefit from this research when dealing with NZ patients suspected of colorectal cancer.


Mary's appointment
Doctor, I have abdominal cramps and ... often constipation
How likely is it that Mary has CRC?
Her generation has high risk of CRC
I will refer you to a gastroenterologist for further investigation
picture from lipsig.com

References

Adelstein, B.-A., Irwig, L., Macaskill, P., Turner, R., Chan, S., and Katelaris, P. (2010). Who needs colonoscopy to identify colorectal cancer? Bowel symptoms do not add substantially to age and other medical history. *Alimentary Pharmacology & Therapeutics*, 32(2):270–281.

Adelstein, B.-A., Macaskill, P., Turner, R. M., Katelaris, P. H., and Irwig, L. (2011). The value of age and medical history for predicting colorectal cancer and adenomas in people referred for colonoscopy. *BMC Gastroenterology*, 11(1):1–10.

Ahlqvist, E., Storm, P., Käräjämäki, A., Martinell, M., Dorkhan, M., Carlsson, A., Vikman, P., Prasad, R. B., Aly, D. M., Almgren, P., et al. (2018). Novel subgroups of adult-onset diabetes and their association with outcomes: a data-driven cluster analysis of six variables. *The Lancet Diabetes & Endocrinology*, 6(5):361–369.

Akobeng, A. K. (2007). Understanding diagnostic tests 3: receiver operating characteristic curves. *Acta Paediatrica*, 96(5):644–647.

Alexiusdottir, K. K., Möller, P. H., Snaebjornsson, P., Jonasson, L., Olafsdottir, E. J., Björnsson, E. S., Tryggvadottir, L., and Jonasson, J. G. (2012). Association of symptoms of colon cancer patients with tumor location and TNM tumor stage. *Scandinavian Journal of Gastroenterology*, 47(7):795–801.

American Diabetes Association (2013). Diagnosis and classification of diabetes mellitus. *Diabetes Care*, 36(S1):S67–S74.

Andersen, B. L., Cacioppo, J. T., and Roberts, D. C. (1995). Delay in seeking a cancer diagnosis: delay stages and psychophysiological comparison processes. *British Journal of Social Psychology*, 34(1):33–52.

Araghi, M., Soerjomataram, I., Bardot, A., Ferlay, J., Cabasag, C. J., Morrison, D. S., De, P., Tervonen, H., Walsh, P. M., Bucher, O., et al. (2019). Changes in colorectal cancer incidence in seven high-income countries: a population-based study. *The Lancet Gastroenterology & Hepatology*, 4(7):511–518.

Arnold, M., Rutherford, M. J., Bardot, A., Ferlay, J., Andersson, T. M., Myklebust, T. Å., Tervonen, H., Thursfield, V., Ransom, D., Shack, L., et al. (2019). Progress in cancer survival, mortality, and incidence in seven high-income countries 1995–2014 (ICBP SURVMARK-2): a population-based study. *The Lancet Oncology*, 20(11):1493–1505.

Arnold, M., Sierra, M. S., Laversanne, M., Soerjomataram, I., Jemal, A., and Bray, F. (2017). Global patterns and trends in colorectal cancer incidence and mortality. *Gut*, 66(4):683–691.

Astin, M., Griffin, T., Neal, R. D., Rose, P., and Hamilton, W. (2011). The diagnostic value of symptoms for colorectal cancer in primary care: a systematic review. *British Journal of General Practice*, 61(586):231–243.

Atkinson, M. A., Eisenbarth, G. S., and Michels, A. W. (2014). Type 1 diabetes. *The Lancet*, 383(9911):69–82.

Atlantis, E., Joshy, G., Williams, M., and Simmons, D. (2017). Diabetes among Māori and other ethnic groups in New Zealand. In *Diabetes mellitus in developing countries and underserved communities*, pages 165–190. Springer.

Augustin, N. H., Sauleau, E.-A., and Wood, S. N. (2012). On quantile quantile plots for generalized linear models. *Computational Statistics & Data Analysis*, 56(8):2404–2409.

Aye, P. S., Elwood, J. M., and Stevanovic, V. (2014). Comparison of cancer survival in New Zealand and Australia, 2006–2010. *The New Zealand Medical Journal (Online)*, 127(1407):14–26.

Bagshaw, P. and Ding, S. (2019). Assessment of Diagnostic and Treatment Times for Endoscopic Cases for Southern District Health Board. `https://www.southernhealth.nz/sites/default/files/2019-07/SDHB%20Endoscopy%20Cases%20Report%20Final%20-%20redacted.pdf`.

Ballotari, P., Vicentini, M., Manicardi, V., Gallo, M., Ranieri, S. C., Greci, M., and Rossi, P. G. (2017). Diabetes and risk of cancer incidence: results from a population-based cohort study in northern Italy. *BMC Cancer*, 17(1):1–8.

Banaszkiewicz, Z., Tojek, K., Jarmocik, P., Frasz, J., and Jawien, A. (2009). Clinical symptoms of colorectal cancer-a retrospective study. *Współczesna Onkologia*, 13(1):34–40.

Beal, E. W., Tumin, D., Moris, D., Zhang, X.-F., Chakedis, J., Dilhoff, M., Schmidt, C. M., and Pawlik, T. M. (2018). Cohort contributions to trends in the incidence and mortality of intrahepatic cholangiocarcinoma. *Hepatobiliary Surgery and Nutrition*, 7(4):270–276.

Bell, A. and Jones, K. (2013). The impossibility of separating age, period and cohort effects. *Social Science & Medicine*, 93(2013):163–165.

Bellentani, S., Baldoni, P., Petrella, S., Tata, C., Armocida, C., Marchegiano, P., Saccoccio, G., Manenti, F., and Group, L. I. S. (1990). A simple score for the identification of patients at high risk of organic diseases of the colon in the family doctor consulting room. *Family Practice*, 7(4):307–312.

Berster, J. M. and Goke, P. D. B. (2008). Type 2 diabetes mellitus as risk factor for colorectal cancer. *Archives of Physiology and Biochemistry*, 114(1):84–98.

Best Practice Advocacy Centre New Zealand (2018). A rising tide of type 2 diabetes in younger people: what can primary care do? `https://bpac.org.nz/2018/diabetes.aspx`.

Bjerregaard, N. C., Tøttrup, A., Sørensen, H. T., and Laurberg, S. (2007). Diagnostic value of self-reported symptoms in Danish outpatients referred with symptoms consistent with colorectal cancer. *Colorectal Disease*, 9(5):443–451.

Blakely, T., Shaw, C., Atkinson, J., Cunningham, R., and Sarfati, D. (2011). Social inequalities or inequities in cancer incidence? Repeated census-cancer cohort studies, New Zealand 1981–1986 to 2001–2004. *Cancer Causes & Control*, 22(9):1307–1318.

Boniol, M. and Heanue, M. (2007). Age-standardisation and denominators. In *Cancer Incidence in five continents*, volume 9, pages 99–101. International Agency for Research on Cancer Scientific Publications.

Bossuyt, P. M., Reitsma, J. B., Bruns, D. E., Gatsonis, C. A., Glasziou, P. P., Irwig, L. M., Moher, D., Rennie, D., De Vet, H. C., and Lijmer, J. G. (2003). The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. *Annals of Internal Medicine*, 138(1):W1–12.

Boyd, J. H., Ferrante, A. M., Irvine, K., Smith, M., Moore, E., Brown, A., and Randall, S. M. (2017). Understanding the origins of record linkage errors and how they affect research outcomes. *Australian and New Zealand Journal of Public Health*, 41(2):215–215.

Boyd, M., Atkinson, J., and Blakely, T. (2016). Ethnic counts on mortality, New Zealand Cancer Registry and census data: 2006–2011. *The New Zealand Medical Journal*, 129(1429):22–39.

Boyle, T., Keegel, T., Bull, F., Heyworth, J., and Fritschi, L. (2012). Physical activity and risks of proximal and distal colon cancers: a systematic review and meta-analysis. *Journal of the National Cancer Institute*, 104(20):1548–1561.

Bray, F. and Møller, B. (2006). Predicting the future burden of cancer. *Nature Reviews Cancer*, 6(1):63–74.

Bray, F. I. and Weiderpass, E. (2010). Lung cancer mortality trends in 36 European countries: secular trends and birth cohort patterns by sex and region 1970–2007. *International Journal of Cancer*, 126(6):1454–1466.

Brenner, D. R., Ruan, Y., Shaw, E., De, P., Heitman, S. J., and Hilsden, R. J. (2017). Increasing colorectal cancer incidence trends among younger adults in Canada. *Preventive Medicine*, 105(2017):345–349.

Buchner, F., Wasem, J., and Schillo, S. (2017). Regression trees identify relevant interactions: can this improve the predictive performance of risk adjustment? *Health Economics*, 26(1):74–85.

Buchwald, P., Hall, C., Davidson, C., Dixon, L., Dobbs, B., Robinson, B., and Frizelle, F. (2018). Improved survival for rectal cancer compared to colon cancer: the four cohort study. *ANZ Journal of Surgery*, 88(3):E114–E117.

Bufill, J. A. (1990). Colorectal cancer: evidence for distinct genetic categories based on proximal or distal tumor location. *Annals of Internal Medicine*, 113(10):779–788.

Burkitt, D. P. (1971). Epidemiology of cancer of the colon and rectum. *Cancer*, 28(1):3–13.

But, A., De Bruin, M. L., Bazelier, M. T., Hjellvik, V., Andersen, M., Auvinen, A., Starup-Linde, J., Schmidt, M. K., Furu, K., de Vries, F., et al. (2017). Cancer risk among insulin users: comparing analogues with human insulin in the CARING five-country cohort study. *Diabetologia*, 60(9):1691–1703.

But, A. et al. (2018). Mitigating bias and dealing with multiple time scales in cohort studies: Studying medications and complications of diabetes. *Dissertationes Scholae Doctoralis Ad Sanitatem Investigandam Universitatis Helsinkiensis*.

Butterworth, A. S., Higgins, J. P., and Pharoah, P. (2006). Relative and absolute risk of colorectal cancer for individuals with a family history: a meta-analysis. *European Journal of Cancer*, 42(2):216–227.

Cameron, C., Samaranayaka, A., and Turner, R. M. (2020). P values: what is their significance? *New Zealand Medical Student Journal*, 11(31):48–49.

Camp, N. J. and Slattery, M. L. (2002). Classification tree analysis: a statistical tool to investigate risk factor interactions with an example for colon cancer (United States). *Cancer Causes & Control*, 13(9):813–823.

Cancer Research UK (2018). Bowel cancer incidence statistics. `http://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/bowel-cancer/incidence`.

Cancer Society NZ (2019). Bowel cancer report shows worrying trend for young people. `https://wellington.cancernz.org.nz/about-us/news-and-media/news/bowel-cancer-report-shows-worrying-trend-for-young-people/`.

Canterbury DHB (2019). Canterbury DHB Information for the Ministry of Health. `http://www.cdhb.health.nz/wp-content/uploads/c4c99575-cdhb-10127-priority-spending-and-cancer-treatment.pdf`.

Carstensen, B. (2007). Age–period–cohort models for the Lexis diagram. *Statistics in Medicine*, 26(15):3018–3045.

Carstensen, B. (2011). Statistical Analysis in the Lexis Diagram: Age-Period-Cohort models. `http://www.bendixcarstensen.com/APC/Lisboa-2011/pracs.pdf`.

Carstensen, B. (2019). Analyses based on the reconstructed Danish Diabetes Register. `http://bendixcarstensen.com/DMreg/NewAna.pdf`.

Carstensen, B., Jørgensen, M. E., and Friis, S. (2014). The Epidemiology of Diabetes and Cancer. *Current Diabetes Report*, 14(10):1–8.

Carstensen, B., Plummer, M., Laara, E., and Hills, M. (2019). Epi: a package for statistical analysis in epidemiology. R package version 2.40. `https://cran.r-project.org/web/packages/Epi/index.html`.

Carstensen, B., Read, S. H., Friis, S., Sund, R., Keskimäki, I., Svensson, A.-M., Ljung, R., Wild, S. H., Kerssens, J. J., Harding, J. L., et al. (2016). Cancer incidence in persons with type 1 diabetes: a five-country study of 9,000 cancers in type 1 diabetic individuals. *Diabetologia*, 59(5):980–988.

Carstensen, B., Witte, D., and Friis, S. (2012). Cancer occurrence in Danish diabetic patients: duration and insulin effects. *Diabetologia*, 55(4):948–958.

Cavestro, G. M., Mannucci, A., Zuppardo, R. A., Di Leo, M., Stoffel, E., and Tonon, G. (2018). Early onset sporadic colorectal cancer: Worrisome trends and oncogenic features. *Digestive and Liver Disease*, 50(6):521–532.

Chan, A. T. and Giovannucci, E. L. (2010). Primary prevention of colorectal cancer. *Gastroenterology*, 138(6):2029–2043.

Chang, C. and Ulrich, C. (2003). Hyperinsulinaemia and hyperglycaemia: possible risk factors of colorectal cancer among diabetic patients. *Diabetologia*, 46(5):595–607.

Cheah, P. Y. (2009). Recent advances in colorectal cancer genetics and diagnostics. *Critical Reviews in Oncology/Hematology*, 69(1):45–55.

Chernyavskiy, P., Little, M. P., and Rosenberg, P. S. (2018). Correlated Poisson models for age-period-cohort analysis. *Statistics in Medicine*, 37(3):405–424.

Chohan, D., Goodwin, K., Wilkinson, S., Miller, R., and Hall, N. (2005). How has the 'two-week wait' rule affected the presentation of colorectal cancer? *Colorectal Disease*, 7(5):450–453.

Chou, C.-L., Chang, S.-C., Lin, T.-C., Chen, W.-S., Jiang, J.-K., Wang, H.-S., Yang, S.-H., Liang, W.-Y., and Lin, J.-K. (2011). Differences in clinicopathological characteristics of colorectal cancer between younger and elderly patients: an analysis of 322 patients from a single institution. *The American Journal of Surgery*, 202(5):574–582.

Chung, R. Y.-N., Tsoi, K. K., Kyaw, M. H., Lui, A. R., Lai, F. T., and Sung, J. J.-Y. (2019). A population-based age-period-cohort study of colorectal cancer incidence comparing Asia against the West. *Cancer Epidemiology*, 59(2019):29–36.

Clayton, D. and Schifflers, E. (1987a). Models for temporal variation in cancer rates. I: age–period and age–cohort models. *Statistics in Medicine*, 6(4):449–467.

Clayton, D. and Schifflers, E. (1987b). Models for temporal variation in cancer rates. II: age–period–cohort models. *Statistics in Medicine*, 6(4):469–481.

Colhoun, H. et al. (2009). Use of insulin glargine and cancer incidence in Scotland: a study from the Scottish Diabetes Research Network Epidemiology Group. *Diabetologia*, 52(9):1755–1765.

Coppell, K. J., Mann, J. I., Williams, S. M., Jo, E., Drury, P. L., Miller, J. C., and Parnell, W. R. (2013). Prevalence of diagnosed and undiagnosed diabetes and prediabetes in New Zealand: findings from the 2008/09 Adult Nutrition Survey. *The New Zealand Medical Journal*, 126(1370):23–42.

Cormack, D., Reid, P., and Kukutai, T. (2019). Indigenous data and health: Critical approaches to 'race'/ethnicity and Indigenous data governance. *Public Health*, 172(2019):116–118.

Cormack, D., Robson, B., Purdie, G., Ratima, M., and Brown, R. (2005). Access to cancer services for Maori. `https://www.health.govt.nz/publication/access-cancer-services-maori`.

Cox, B. (2016). Flexible sigmoidoscopy is the best approach for a national bowel screening programme. *The New Zealand Medical Journal*, 129(1430):14–7.

Cox, B. and Little, J. (1992). Reduced risk of colorectal cancer among recent generations in New Zealand. *British Journal of Cancer*, 66(2):386–390.

Cunningham, R., Sarfati, D., Hill, S., and Kenwright, D. (2008). An audit of colon cancer data on the New Zealand Cancer Registry. *The New Zealand Medical Journal (Online)*, 121(1279):46–56.

Currie, C., Poole, C., and Gale, E. (2009). The influence of glucose-lowering therapies on cancer risk in type 2 diabetes. *Diabetologia*, 52(9):1766–1777.

Dankner, R., Boffetta, P., Balicer, R. D., Boker, L. K., Sadeh, M., Berlin, A., Olmer, L., Goldfracht, M., and Freedman, L. S. (2016). Time-dependent risk of cancer after a diabetes diagnosis in a cohort of 2.3 million adults. *American Journal of Epidemiology*, 183(12):1098–1106.

De Bruijn, K. M., Ruiter, R., de Keyser, C. E., Hofman, A., Stricker, B. H., and van Eijck, C. H. (2014). Detection bias may be the main cause of increased cancer incidence among diabetics: results from the Rotterdam Study. *European Journal of Cancer*, 50(14):2449–2455.

de Jong, R., Burden, A., de Kort, S., van Herk-Sukel, M., Vissers, P., Janssen, P., Haak, H., Masclee, A., de Vries, F., and Janssen-Heijnen, M. (2017). Impact of detection bias on the risk of gastrointestinal cancer and its subsites in type 2 diabetes mellitus. *European Journal of Cancer*, 79(2017):61–71.

de Kort, S., Masclee, A. A., Sanduleanu, S., Weijenberg, M. P., van Herk-Sukel, M. P., Oldenhof, N. J., Van Den Bergh, J. P., Haak, H. R., and Janssen-Heijnen, M. L. (2017). Higher risk of colorectal cancer in patients with newly diagnosed diabetes mellitus before the age of colorectal cancer screening initiation. *Scientific Reports*, 7(1):1–8.

de Kort, S., Simons, C., van den Brandt, P. A., Goldbohm, R. A., Arts, I. C., de Bruine, A. P., Janssen-Heijnen, M. L., Sanduleanu, S., Masclee, A. A., and Weijenberg, M. P. (2016). Diabetes mellitus type 2 and subsite-specific colorectal cancer risk in men and women: results from the Netherlands Cohort Study on diet and cancer. *European Journal of Gastroenterology & Hepatology*, 28(8):896–903.

Dharwadkar, P., Greenan, G., Singal, A. G., and Murphy, C. C. (2019). Is Colorectal Cancer in Patients Younger Than 50 Years of Age the Same Disease as in Older Patients? *Clinical Gastroenterology and Hepatology*, 19(1):192–194.

Didham, R. and Callister, P. (2012). The effect of ethnic prioritisation on ethnic health analysis: a research note. *The New Zealand Medical Journal*, 125(1359):58–66.

Dobson, A., Hockey, R., Chan, H.-W., and Mishra, G. (2020). Flexible age-period-cohort modelling illustrated using obesity prevalence data. *BMC Medical Research Methodology*, 20(1):1–9.

Dobson, A. J. and Barnett, A. (2008). An Introduction to Generalized Linear Models. In *An Introduction to Generalized Linear Models*. CRC Press.

Dommett, R., Redaniel, M., Stevens, M., Hamilton, W., and Martin, R. (2013). Features of cancer in teenagers and young adults in primary care: a population-based nested case–control study. *British Journal of Cancer*, 108(11):2329–2333.

Douaiher, J., Ravipati, A., Grams, B., Chowdhury, S., Alatise, O., and Are, C. (2017). Colorectal cancer - global burden, trends, and geographical variations. *Journal of Surgical Oncology*, 115(5):619–630.

Dubrow, R., Bernstein, J., and Holford, T. R. (1993). Age-period-cohort modelling of large-bowel-cancer incidence by anatomic sub-site and sex in connecticut. *International Journal of Cancer*, 53(6):907–913.

Dubrow, R., Johansen, C., Skov, T., and Holford, T. R. (1994). Age-period-cohort modelling of large-bowel-cancer incidence by anatomic sub-site and sex in Denmark. *International Journal of Cancer*, 58(3):324–329.

Durrleman, S. and Simon, R. (1989). Flexible regression models with cubic splines. *Statistics in Medicine*, 8(5):551–561.

Elstein, A. S. and Schwarz, A. (2002). Clinical problem solving and diagnostic decision making: selective review of the cognitive literature. *British Medical Journal*, 324(7339):729–732.

Elwood, J. M., Aye, P. S., and Tin Tin, S. (2016). Increasing Disadvantages in Cancer Survival in New Zealand Compared to Australia, between 2000-05 and 2006-10. *PLOS ONE*, 11(3):1–14.

Enders, C. K. (2010). *Applied missing data analysis*. Guilford press.

Ewing, M. (2018). Identification and early detection of cancer patients in primary care. `http://hdl.handle.net/2077/55379`.

Ewing, M., Naredi, P., Zhang, C., and Månsson, J. (2016). Identification of patients with non-metastatic colorectal cancer in primary care: a case-control study. *British Journal of General Practice*, 66(653):e880–e886.

Favoriti, P., Carbone, G., Greco, M., Pirozzi, F., Pirozzi, R. E. M., and Corcione, F. (2016). Worldwide burden of colorectal cancer: a review. *Updates in Surgery*, 68(1):7–11.

Feletto, E., Yu, X. Q., Lew, J.-B., St John, D. J. B., Jenkins, M. A., Macrae, F. A., Mahady, S. E., and Canfell, K. (2019). Trends in colon and rectal cancer incidence in

Australia from 1982 to 2014: analysis of data on over 375,000 cases. *Cancer Epidemiology and Prevention Biomarkers*, 28(1):83–90.

Firth, M. J., Sharples, K. J., Hinder, V. A., Macapagal, J., Sarfati, D., Derret, S. L., Hill, A. G., Brown, C., Reid, P. M., Lawrenson, R., et al. (2016). Methods of a national colorectal cancer cohort study: the PIPER Project. *New Zealand Medical Journal*, 129(1440):25–36.

Fletcher, R. H. (2009). The diagnosis of colorectal cancer in patients with symptoms: finding a needle in a haystack. *BMC Medicine*, 7(1):1–3.

Ford, A. C., Van Zanten, S. V., Rodgers, C. C., Talley, N. J., Vakil, N., and Moayyedi, P. (2008). Diagnostic utility of alarm features for colorectal cancer: systematic review and meta-analysis. *Gut*, 57(11):1545–1553.

Ford, E., Carroll, J. A., Smith, H. E., Scott, D., and Cassell, J. A. (2016). Extracting information from the text of electronic medical records to improve case detection: a systematic review. *Journal of the American Medical Informatics Association*, 23(5):1007–1015.

Fosse, E. and Winship, C. (2019). Analyzing Age-Period-Cohort Data: A Review and Critique. *Annual Review of Sociology*, 45:467–492.

Gaertner, W. B., Kwaan, M. R., Madoff, R. D., and Melton, G. B. (2015). Rectal cancer: An evidence-based update for primary care providers. *World Journal of Gastroenterology: WJG*, 21(25):7659–7671.

Gandhi, J., Davidson, C., Hall, C., Pearson, J., Eglinton, T., Wakeman, C., and Frizelle, F. (2017). Population-based study demonstrating an increase in colorectal cancer in young patients. *British Journal of Surgery*, 104(8):1063–1068.

Gandomani, H. S., Aghajani, M., Mohammadian-Hafshejani, A., Tarazoj, A. A., Pouyesh,

V., Salehiniya, H., et al. (2017). Colorectal cancer in the world: incidence, mortality and risk factors. *Biomedical Research and Therapy*, 4(10):1656–1675.

Gatta, G., Ciccolallo, L., Capocaccia, R., Coleman, M., Hakulinen, T., Møller, H., Berrino, F., Group, E. W., et al. (2003). Differences in colorectal cancer survival between European and US populations: the importance of sub-site and morphology. *European Journal of Cancer*, 39(15):2214–2222.

Gavrielov-Yusim, N. and Friger, M. (2014). Use of administrative medical databases in population-based research. *Journal of Epidemiology and Community Health*, 68(3):283–287.

Gillies, R. J., Pilot, C., Marunaka, Y., and Fais, S. (2019). Targeting acidity in cancer and diabetes. *Biochimica et Biophysica Acta (BBA)-Reviews on Cancer*, 1871(2):273–280.

Gini, A., Bidoli, E., Zanier, L., Clagnan, E., Zanette, G., Gobbato, M., De Paoli, P., and Serraino, D. (2016). Cancer among patients with type 2 diabetes mellitus: a population-based cohort study in northeastern Italy. *Cancer Epidemiology*, 41(2016):80–87.

Giouleme, O., Diamantidis, M. D., and Katsaros, M. G. (2011). Is diabetes a causal agent for colorectal cancer? Pathophysiological and molecular mechanisms. *World Journal of Gastroenterology: WJG*, 17(4):444–448.

Giovannucci, E. (2002). Modifiable risk factors for colon cancer. *Gastroenterology Clinics of North America*, 31(4):925–943.

Glaser, S. L., Clarke, C. A., Gomez, S. L., O?Malley, C. D., Purdie, D. M., and West, D. W. (2005). Cancer surveillance research: a vital subdiscipline of cancer epidemiology. *Cancer Causes & Control*, 16(9):1009–1019.

Gonzalez, E. C., Roetzheim, R. G., Ferrante, J. M., and Campbell, R. (2001). Predictors of proximal vs. distal colorectal cancers. *Diseases of the Colon & Rectum*, 44(2):251–258.

González, N., Prieto, I., del Puerto-Nevado, L., Portal-Nuñez, S., Ardura, J. A., Corton, M., Fernández-Fernández, B., Aguilera, O., Gomez-Guerrero, C., Mas, S., et al. (2017). 2017 update on the relationship between diabetes and colorectal cancer: epidemiology, potential molecular mechanisms and therapeutic implications. *Oncotarget*, 8(11):18456.

Greenland, S. (1989). Modeling and variable selection in epidemiologic analysis. *American Journal of Public Health*, 79(3):340–349.

Grimes, D. A. and Schulz, K. F. (2002). Bias and causal associations in observational research. *The Lancet*, 359(9302):248–252.

Gurney, J. K., Sarfati, D., Lawrence, B., Jackson, C., Findlay, M., and McPherson, K. (2020). Cancer research in the New Zealand context: Challenges and advantages. *Journal of Cancer Policy*, 23(2020):100204.

Haggar, F. A. and Boushey, R. P. (2009). Colorectal cancer epidemiology: incidence, mortality, survival, and risk factors. *Clinics in Colon and Rectal Surgery*, 22(04):191–197.

Half, E. and Arber, N. (2009). Colon cancer: preventive agents and the present status of chemoprevention. *Expert Opinion on Pharmacotherapy*, 10(2):211–219.

Halso- och Sjukvard (2012). Cancer Incidence in Sweden 2011. `https://www.socialstyrelsen.se/globalassets/sharepoint-dokument/artikelkatalog/statistik/2012-12-19.pdf`.

Hamilton, W. (2009a). Five misconceptions in cancer diagnosis. *British Journal of General Practice*, 59(563):441–447.

Hamilton, W. (2009b). The CAPER studies: five case-control studies aimed at identifying and quantifying the risk of cancer in symptomatic primary care patients. *British journal of cancer*, 101(2):S80–S86.

Hamilton, W., Lancashire, R., Sharp, D., Peters, T. J., Cheng, K., and Marshall, T. (2009). The risk of colorectal cancer with symptoms at different ages and between the sexes: a case-control study. *BMC Medicine*, 7(1):1–9.

Hamilton, W., Round, A., Sharp, D., and Peters, T. (2005). Clinical features of colorectal cancer before diagnosis: a population-based case–control study. *British Journal of Cancer*, 93(4):399–405.

Hamilton, W. and Sharp, D. (2004). Diagnosis of colorectal cancer in primary care: the evidence base for guidelines. *Family Practice*, 21(1):99–106.

Harding, J. L., Shaw, J. E., Peeters, A., Cartensen, B., and Magliano, D. J. (2015). Cancer risk among people with type 1 and type 2 diabetes: disentangling true associations, detection bias, and reverse causation. *Diabetes Care*, 38(2):264–270.

He, J., Stram, D., Kolonel, L., Henderson, B., Le Marchand, L., and Haiman, C. (2010). The association of diabetes with colorectal cancer risk: the Multiethnic Cohort. *British Journal of Cancer*, 103(1):120–126.

Hemkens, L. G., Grouven, U., Bender, R., Günster, C., Gutschmidt, S., Selke, G. W., and Sawicki, P. T. (2009). Risk of malignancies in patients with diabetes treated with human insulin or insulin analogues: a cohort study. *Diabetologia*, 52(9):1732–1744.

Hernandez-Diaz, S. and Adami, H.-O. (2010). Diabetes therapy and cancer risk: causal effects and other plausible explanations. *Diabetologia*, 53(5):802–808.

Heuer, C. (1997). Modeling of time trends and interactions in vital rates using restricted regression splines. *Biometrics*, 53(1997):161–177.

Hill, S., Sarfati, D., Blakely, T., Robson, B., Purdie, G., Chen, J., Dennett, E., Cormack, D., Cunningham, R., Dew, K., et al. (2010). Survival disparities in Indigenous and non-Indigenous New Zealanders with colon cancer: the role of patient comorbidity,

treatment and health service factors. *Journal of Epidemiology & Community Health*, 64(2):117–123.

Hippisley-Cox, J. and Coupland, C. (2012). Identifying patients with suspected colorectal cancer in primary care: derivation and validation of an algorithm. *British Journal of General Practice*, 62(594):e29–e37.

Hippisley-Cox, J. and Coupland, C. (2013). Symptoms and risk factors to identify women with suspected cancer in primary care: derivation and validation of an algorithm. *British Journal of General Practice*, 63(606):e11–e21.

Hirsz, M., Hunt, L., Chepulis, L., and Mayo, M. (2019). Can we select patients for colonoscopy more accurately? In *Proceedings of the Waikato Clinical Campus Biannual Research Seminar*.

Hofseth, L. J., Hebert, J. R., Chanda, A., Chen, H., Love, B. L., Pena, M. M., Murphy, E. A., Sajish, M., Sheth, A., Buckhaults, P. J., et al. (2020). Early-onset colorectal cancer: initial clues and current views. *Nature Reviews Gastroenterology & Hepatology*, 17(6):352–364.

Holford, T. R. (1983). The estimation of age, period and cohort effects for vital rates. *Biometrics*, 39(1983):311–324.

Holford, T. R. (1991). Understanding the effects of age, period, and cohort on incidence and mortality rates. *Annual Review of Public Health*, 12(1):425–457.

Holford, T. R., Chen, H.-S., Annett, D., Krapcho, M., Dorogaeva, A., and Feuer, E. J. (2019). CP* Trends: An Online Tool for Comparing Cohort and Period Trends Across Cancer Sites. *American Journal of Epidemiology*, 188(7):1361–1370.

Hoover, R. N., Hyer, M., Pfeiffer, R. M., Adam, E., Bond, B., Cheville, A. L., Colton, T., Hartge, P., Hatch, E. E., Herbst, A. L., et al. (2011). Adverse health outcomes

in women exposed in utero to diethylstilbestrol. *New England Journal of Medicine*, 365(14):1304–1314.

Hosmer Jr, D. W., Lemeshow, S., and Sturdivant, R. X. (2013). *Applied logistic regression*, volume 398. John Wiley & Sons.

Hothorn, T. and Everitt, B. S. (2014). *A handbook of statistical analyses using R*. CRC press.

Hsiang, J. C., Bai, W., and Lal, D. (2013). Symptom presentations and other characteristics of colorectal cancer patients and the diagnostic performance of the Auckland Regional Grading Criteria for Suspected Colorectal Cancer in the South Auckland population. *The New Zealand Medical Journal*, 126(1382):95–107.

International Agency for Research on Cancer (2012). GLOBOCAN 2012: estimated cancer incidence, mortality and prevalence worldwide in 2012. `http://globocan.iarc.fr/Pages/summary_table_pop_prev_sel.aspx`.

International Agency for Research on Cancer (2020). Globocan 2020: Colorectal Cancer. `http://gco.iarc.fr/today/data/factsheets/cancers/10_8_9-Colorectum-fact-sheet.pdf`.

Jalal, H. and Burke, D. S. (2019). Hexamaps for Visualizing Age-Period-Cohort Data Trends. *BMJ MedRxiv*, 1(2019):1–11.

Jeffreys, M., Stevanovic, V., Tobias, M., Lewis, C., Ellison-Loschmann, L., Pearce, N., and Blakely, T. (2005). Ethnic inequalities in cancer survival in New Zealand: linkage study. *American Journal of Public Health*, 95(5):834–837.

Jellema, P., van der Windt, D., Bruinvels, D., Mallen, C., van Weyenberg, S., Mulder, C., and de Vet, H. (2010). Value of symptoms and additional diagnostic tests for colorectal cancer in primary care: systematic review and meta-analysis. *British Medical Journal*, 2010(2010):340.

Jemal, A., Travis, W. D., Tarone, R. E., Travis, L., and Devesa, S. S. (2003). Lung cancer rates convergence in young men and women in the United States: analysis by birth cohort and histologic type. *International Journal of Cancer*, 105(1):101–107.

Jiang, Y., Ben, Q., Shen, H., Lu, W., Zhang, Y., and Zhu, J. (2011). Diabetes mellitus and incidence and mortality of colorectal cancer: a systematic review and meta-analysis of cohort studies. *European Journal of Epidemiology*, 26(11):863–876.

Jo, C., Wright, C., Dawson, A., Orr-Walker, B., and Drury, P. (2010). The development and validation of a 'Virtual Diabetes Registry' (VDR) for monitoring diabetes prevalence and the quality of diabetes care in New Zealand. In *The 8th International Diabetes Federation Western Pacific Region Congress*.

Johnson, C. M., Wei, C., Ensor, J. E., Smolenski, D. J., Amos, C. I., Levin, B., and Berry, D. A. (2013). Meta-analyses of colorectal cancer risk factors. *Cancer Causes & Control*, 24(6):1207–1222.

Johnson, J., Bowker, S., Richardson, K., and Marra, C. (2011). Time-varying incidence of cancer after the onset of type 2 diabetes: evidence of potential detection bias. *Diabetologia*, 54(9):2263–2271.

Johnson, J., Carstensen, B., Witte, D., Bowker, S., Lipscombe, L., Renehan, A., Diabetes, Consortium, C. R., et al. (2012). Diabetes and cancer (1): evaluating the temporal relationship between type 2 diabetes and cancer incidence. *Diabetologia*, 55(6):1607–1618.

Jonasson, J., Ljung, R., Talbäck, M., Haglund, B., Gudbjörnsdòttir, S., and Steineck, G. (2009). Insulin glargine use and short-term incidence of malignancies: a population-based follow-up study in Sweden. *Diabetologia*, 52(9):1745–1754.

Jones, R. (2008). Primary care research and clinical practice: gastroenterology. *Postgraduate Medical Journal*, 84(995):454–458.

Jones, R., Latinovic, R., Charlton, J., and Gulliford, M. C. (2007). Alarm symptoms in early diagnosis of cancer in primary care: cohort study using General Practice Research Database. *British Medical Journal*, 334(7602):1040.

Jones, R., Rubin, G., and Hungin, P. (2001). Is the two week rule for cancer referrals working?: Not too well. *BMJ: British Medical Journal*, 322(7302):1555.

Juillerat, P., Peytremann-Bridevaux, I., Vader, J.-P., Arditi, C., Filliettaz, S. S., Dubois, R., Gonvers, J.-J., Froehlich, F., Burnand, B., and Pittet, V. (2009). Appropriateness of colonoscopy in Europe (EPAGE II)–presentation of methodology, general results, and analysis of complications. *Endoscopy*, 41(03):240–246.

Kaminski, M. F., Polkowski, M., Kraszewska, E., Rupinski, M., Butruk, E., and Regula, J. (2014). A score to estimate the likelihood of detecting advanced colorectal neoplasia at colonoscopy. *Gut*, 63(7):1112–1119.

Keiding, N. (1990). Statistical inference in the Lexis diagram. *Philosophical Transactions of the Royal Society of London. Series A: Physical and Engineering Sciences*, 332(1627):487–509.

Kelty, E., Ward, S. V., Cadby, G., McCarthy, N. S., O?Leary, P., Moses, E. K., Ee, H. C., and Preen, D. B. (2019). Familial and non-familial risk factors associated with incidence of colorectal cancer in young and middle-aged persons in Western Australia. *Cancer Epidemiology*, 62(2019):101591.

Keum, N. and Giovannucci, E. (2019). Global burden of colorectal cancer: emerging trends, risk factors and prevention strategies. *Nature Reviews Gastroenterology & Hepatology*, 16(12):713–732.

Keyes, K. M., Utz, R. L., Robinson, W., and Li, G. (2010). What is a cohort effect? Comparison of three statistical methods for modelling cohort effects in obesity prevalence in the United States, 1971–2006. *Social Science & Medicine*, 70(7):1100–1108.

Khalid-de Bakker, C., Jonkers, D., Smits, K., Mesters, I., Masclee, A., and Stockbrügger, R. (2011). Participation in colorectal cancer screening trials after first-time invitation: a systematic review. *Endoscopy*, 43(12):1059.

Kinar, Y., Kalkstein, N., Akiva, P., Levin, B., Half, E. E., Goldshtein, I., Chodick, G., and Shalev, V. (2016). Development and validation of a predictive model for detection of colorectal cancer in primary care by analysis of complete blood counts: a binational retrospective study. *Journal of the American Medical Informatics Association*, 23(5):879–890.

Koeling, R., Tate, A. R., and Carroll, J. A. (2011). Automatically estimating the incidence of symptoms recorded in GP free text notes. In *Proceedings of the first international workshop on managing interoperability and complexity in health systems*, pages 43–50. ACM.

Kolligs, F. T. (2016). Diagnostics and epidemiology of colorectal cancer. *Visceral Medicine*, 32(3):158–164.

Langenbach, M. R., Schmidt, J., Neumann, J., and Zirngibl, H. (2003). Delay in treatment of colorectal cancer: multifactorial problem. *World Journal of Surgery*, 27(3):304–308.

Larsen, I. K. and Bray, F. (2010). Trends in colorectal cancer incidence in Norway 1962–2006: an interpretation of the temporal patterns by anatomic subsite. *International Journal of Cancer*, 126(3):721–732.

Larsson, S. C., Orsini, N., and Wolk, A. (2005). Diabetes mellitus and risk of colorectal cancer: a meta-analysis. *Journal of the National Cancer Institute*, 97(22):1679–1687.

Last, J. M., Harris, S. S., Thuriaux, M. C., and Spasoff, R. A. (2001). *A dictionary of epidemiology*. International Epidemiological Association, Inc.

Laugesen, M. (2000). The New Zealand food supply and diet-trends 1961-95 and comparison with other OECD countries. *New Zealand Medical Journal*, 113(1114):311–315.

Lawler, M., Naredi, P., Cufer, T., Banks, I., Lievens, Y., Vassal, G., Aapro, M., Sotlar, M. J., Philip, T., Jassem, J., et al. (2019). Moonshot or groundshot: addressing Europe's cancer challenge through a patient-focused, data-enabled lens. *The Lancet Oncology*, 20(11):1482–1485.

Leaman, Aaron (2017). Concerns raised over declined colonoscopy referrals. `https://www.stuff.co.nz/national/health/95332656/concerns-raised-over-declined-colonoscopy-referrals`.

Levin, B., Lieberman, D. A., McFarland, B., Andrews, K. S., Brooks, D., Bond, J., Dash, C., Giardiello, F. M., Glick, S., Johnson, D., et al. (2008). Screening and surveillance for the early detection of colorectal cancer and adenomatous polyps, 2008: a joint guideline from the American Cancer Society, the US Multi-Society Task Force on Colorectal Cancer, and the American College of Radiology. *Gastroenterology*, 134(5):1570–1595.

Limburg, P. J., Anderson, K. E., Johnson, T. W., Jacobs, D. R., Lazovich, D., Hong, C.-P., Nicodemus, K. K., and Folsom, A. R. (2005). Diabetes mellitus and subsite-specific colorectal cancer risks in the Iowa Women's Health Study. *Cancer Epidemiology and Prevention Biomarkers*, 14(1):133–137.

Liu, X., Hemminki, K., Försti, A., Sundquist, K., Sundquist, J., and Ji, J. (2015). Cancer risk in patients with type 2 diabetes mellitus and their relatives. *International Journal of Cancer*, 137(4):903–910.

Lowy, D. R. and Singer, D. S. (2017). Implementing the Cancer Moonshot and beyond. *The Lancet Oncology*, 18(11):e622–e623.

Luo, S., Li, J.-Y., Zhao, L.-N., Yu, T., Zhong, W., Xia, Z.-S., Shan, T.-D., Ouyang, H., Yang, H.-S., and Chen, Q.-K. (2016). Diabetes mellitus increases the risk of colorectal neoplasia: an updated meta-analysis. *Clinics and Research in Hepatology and Gastroenterology*, 40(1):110–123.

Lyratzopoulos, G., Wardle, J., and Rubin, G. (2014). Rethinking diagnostic delay in cancer: how difficult is the diagnosis? *BMJ: British Medical Journal (Online)*, 349(2014):g7400.

Majek, O., Gondos, A., Jansen, L., Emrich, K., Holleczek, B., Katalinic, A., Nennecke, A., Eberle, A., and Brenner, H. (2012). Survival from colorectal cancer in Germany in the early 21st century. *British Journal of Cancer*, 106(11):1875–1880.

Majumdar, S. R., Fletcher, R. H., and Evans, A. T. (1999). How does colorectal cancer present? Symptoms, duration, and clues to location. *The American Journal of Gastroenterology*, 94(10):3039–3045.

Marshall, T., Lancashire, R., Sharp, D., Peters, T. J., Cheng, K., and Hamilton, W. (2011). The diagnostic performance of scoring systems to identify symptomatic colorectal cancer compared to current referral guidance. *Gut*, 60(9):1242–1248.

Mauri, G., Sartore-Bianchi, A., Russo, A.-G., Marsoni, S., Bardelli, A., and Siena, S. (2019). Early-onset colorectal cancer in young individuals. *Molecular Oncology*, 13(2):109–131.

McArdle, C. and Hole, D. (2004). Emergency presentation of colorectal cancer is associated with poor 5-year survival. *British Journal of Surgery*, 91(5):605–609.

McKergow, E., Parkin, L., Barson, D. J., Sharples, K. J., and Wheeler, B. J. (2017). Demographic and regional disparities in insulin pump utilization in a setting of universal funding: a New Zealand nationwide study. *Acta Diabetologica*, 54(1):63–71.

McLeod, M., Harris, R., Crengle, S., Cormack, D., Scott, N., and Robson, B. (2021). Bowel cancer screening age range for māori: what is all the fuss about? *NZM J*, pages 71–77.

Meystre, S. M., Savova, G. K., Kipper-Schuler, K. C., and Hurdle, J. F. (2008). Extracting

information from textual documents in the electronic health record: a review of recent research. *Yearbook of Medical Informatics*, 17(01):128–144.

Mikaeel, R. R., Symonds, E. L., Kimber, J., Smith, E., Horsnell, M., Uylaki, W., Tapia Rico, G., Hewett, P. J., Yong, J., Tonkin, D., et al. (2021). Young-onset colorectal cancer is associated with a personal history of type 2 diabetes. *Asia-Pacific Journal of Clinical Oncology*, 17(1):131–138.

Ministry of Health (2016). Cancer: historical summary 1948-2013. https://www.health.govt.nz/publication/cancer-historical-summary-1948-2013.

Ministry of Health and the NZ Cancer Control Trust (2003). The New Zealand Cancer Control Strategy. https://www.health.govt.nz/system/files/documents/publications/cancercontrolstrategy.pdf.

Ministry of Health NZ (2010). Cancer Projections: Incidence 2004–08 to 2014–18. https://www.health.govt.nz/system/files/documents/publications/cancer-projections-incidence-2004-08-to-2014-18.pdf.

Ministry of Health NZ (2012). New Zealand Primary Care Handbook 2012. https://www.health.govt.nz/system/files/documents/publications/nz-primary-care_handbook_2012.pdf.

Ministry of Health NZ (2016). Final evaluation report of bowel screening pilot. https://www.health.govt.nz/system/files/documents/publications/bowel-screening-pilot-final-evaluation-report-redacted-january2017.docx.

Ministry of Health NZ (2017). National Bowel Screening Programme - Interim Quality Standards. https://www.nsu.govt.nz/system/files/resources/national-bowel-screening-programme-interim-quality-standards-dec17.pdf.

Ministry of Health NZ (2018a). Cancer data and stats. https://www.health.govt.nz/

nz-health-statistics/health-statistics-and-data-sets/cancer-data-and-stats.

Ministry of Health NZ (2018b). National Bowel Screening Programme: Consideration of the potential equity impacts for Māori of the age range for screening. https://www.health.govt.nz/system/files/documents/pages/nbsp-considering-potential-equity-impacts-for-maori.pdf.

Ministry of Health NZ (2018c). Virtual Diabetes Register Guide for Use Statement. Received by email from MoH 25 March 2019.

Ministry of Health NZ (2019a). Mortality 2017 Data Tables. https://www.health.govt.nz/system/files/documents/publications/mort-2017-pub-20191218-final.xlsx.

Ministry of Health NZ (2019b). New Cancer Registrations 2017. https://www.health.govt.nz/system/files/documents/publications/new-cancer-registrations-2017-dec19_0.xlsx.

Ministry of Health NZ (2019c). Referral Criteria for Direct Access Outpatient Colonoscopy or Computed Tomography Colonography. https://www.health.govt.nz/publication/referral-criteria-direct-access-outpatient-colonoscopy-or-computed-tomography-colonography.

Ministry of Health NZ (2019d). Selected Cancers 2015–2017. https://www.health.govt.nz/system/files/documents/publications/selected-cancers-2015-16-17-may2019.xlsx.

Ministry of Health NZ (2019e). Virtual Diabetes Register (VDR). https://www.health.govt.nz/our-work/diseases-and-conditions/diabetes/about-diabetes/virtual-diabetes-register-vdr.

Ministry of Health NZ (2020a). Cancer: Historical summary 1948–2017. https://www.health.govt.nz/publication/cancer-historical-summary-1948-2017.

Ministry of Health NZ (2020b). New Zealand Cancer Registry. https://www.health.govt.nz/nz-health-statistics/national-collections-and-surveys/collections/new-zealand-cancer-registry-nzcr.

Ministry of Health NZ (2020c). Referral of patients with features suggestive of bowel cancer: Ministry of Health guidance. https://bpac.org.nz/2020/docs/bowel-cancer.pdf.

Moons, K. G., Altman, D. G., Reitsma, J. B., Ioannidis, J. P., Macaskill, P., Steyerberg, E. W., Vickers, A. J., Ransohoff, D. F., and Collins, G. S. (2015). Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. *Annals of Internal Medicine*, 162(1):W1–W73.

Moore, S. P., Antoni, S., Colquhoun, A., Healy, B., Ellison-Loschmann, L., Potter, J. D., Garvey, G., and Bray, F. (2015). Cancer incidence in indigenous people in Australia, New Zealand, Canada, and the USA: a comparative population-based study. *The Lancet Oncology*, 16(15):1483–1492.

Murphy, C. C. and Yang, Y. C. (2018). Use of age-period-cohort analysis in cancer epidemiology research. *Current Epidemiology Reports*, 5(4):418–431.

National Collaborating Centre for Cancer (2015). Suspected cancer: recognition and referral. https://www.ncbi.nlm.nih.gov/books/NBK328456.

National Institute for Health and Clinical Excellence (2005). *Referral guidelines for suspected cancer*. National Institute for Health and Clinical Excellence (London).

New Zealand Guidelines Group (2009). Suspected Cancer in Primary Care: Guidelines for investigation, referral and reducing ethnic disparities. https://www.health.govt.

nz/publication/suspected-cancer-primary-care-guidelines-investigation-referral-and-reducing-ethnic-disparities.

Nielsen, H. J., Jakobsen, K. V., Christensen, I. J., and Brünner, N. (2011). Screening for colorectal cancer: possible improvements by risk assessment evaluation? *Scandinavian Journal of Gastroenterology*, 46(11):1283–1294.

Norat, T., Bingham, S., Ferrari, P., Slimani, N., Jenab, M., Mazuir, M., Overvad, K., Olsen, A., Tjønneland, A., Clavel, F., et al. (2005). Meat, fish, and colorectal cancer risk: the European Prospective Investigation into cancer and nutrition. *Journal of the National Cancer Institute*, 97(12):906–916.

O'Connell, J. B., Maggard, M. A., and Ko, C. Y. (2004). Colon cancer survival rates with the new American Joint Committee on Cancer sixth edition staging. *Journal of the National Cancer Institute*, 96(19):1420–1425.

of Health, M. (2014). Screening, Diagnosis and Management of Gestational Diabetes in New Zealand. https://www.health.govt.nz/system/files/documents/publications/screening-diagnosis-management-of-gestational-diabetes-in-nz-clinical-practive-guideline-dec14-v2.docx.

Ogurtsova, K., da Rocha Fernandes, J., Huang, Y., Linnenkamp, U., Guariguata, L., Cho, N. H., Cavan, D., Shaw, J., and Makaroff, L. (2017). IDF Diabetes Atlas: Global estimates for the prevalence of diabetes for 2015 and 2040. *Diabetes Research and Clinical Practice*, 128:40–50.

Ohkuma, T., Peters, S. A., and Woodward, M. (2018). Sex differences in the association between diabetes and cancer: a systematic review and meta-analysis of 121 cohorts including 20 million individuals and one million events. *Diabetologia*, 61(10):2140–2154.

O'Sullivan, J. W., Banerjee, A., Heneghan, C., and Pluddemann, A. (2018). Verification bias. *BMJ Evidence-Based Medicine*, 23(2):54–55.

Overbeek, J. A., Kuiper, J. G., van der Heijden, A. A., Labots, M., Haug, U., Herings, R. M., and Nijpels, G. (2019). Sex-and site-specific differences in colorectal cancer risk among people with type 2 diabetes. *International Journal of Colorectal Disease*, 34(2):269–276.

Pang, Y., Kartsonaki, C., Guo, Y., Chen, Y., Yang, L., Bian, Z., Bragg, F., Millwood, I. Y., Shen, L., Zhou, S., et al. (2018). Diabetes, plasma glucose and incidence of colorectal cancer in Chinese adults: a prospective study of 0.5 million people. *Journal of Epidemiology and Community Health*, 72(10):919–925.

Panzuto, F., Chiriatti, A., Bevilacqua, S., Giovannetti, P., Russo, G., Impinna, S., Pistilli, F., Capurso, G., Annibale, B., Delle Fave, G., et al. (2003). Symptom-based approach to colorectal cancer: survey of primary care physicians in Italy. *Digestive and Liver Disease*, 35(12):869–875.

Parkin, D. M. (2008). The role of cancer registries in cancer control. *International Journal of Clinical Oncology*, 13(2):102–111.

Paternoster, R., Brame, R., Mazerolle, P., and Piquero, A. (1998). Using the correct statistical test for the equality of regression coefficients. *Criminology*, 36(4):859–866.

Patrick, J. and Asgari, P. (2010). Analysing Clinical Notes for Translation Research: Back to the Future. In *Health Information Systems: Concepts, Methodologies, Tools, and Applications*, pages 1954–1975. IGI Global.

Peeters, P. J., Bazelier, M. T., Leufkens, H. G., de Vries, F., and De Bruin, M. L. (2015). The risk of colorectal cancer in patients with type 2 diabetes: associations with treatment stage and obesity. *Diabetes Care*, 38(3):495–502.

Pepe, M. S. et al. (2003). *The statistical evaluation of medical tests for classification and prediction*. Medicine.

Platz, E. A., Willett, W. C., Colditz, G. A., Rimm, E. B., Spiegelman, D., and Giovannucci, E. (2000). Proportion of colon cancer risk that might be preventable in a cohort of middle-aged US men. *Cancer Causes & Control*, 11(7):579–588.

Pocock, S. J. and Smeeth, L. (2009). Insulin glargine and malignancy: an unwarranted alarm. *The Lancet*, 374(9689):511–513.

Ransohoff, D. F. (2005). Colon cancer screening in 2005: status and challenges. *Gastroenterology*, 128(6):1685–1695.

Rawla, P., Sunkara, T., and Barsouk, A. (2019). Epidemiology of colorectal cancer: Incidence, mortality, survival, and risk factors. *Gastroenterology Review*, 14(2):89–103.

Resnik, P., Niv, M., Nossal, M., Kapit, A., and Toren, R. (2008). Communication of clinically relevant information in electronic health records: a comparison between structured data and unrestricted physician language. In *Perspectives in Health Information Management*. CAC Proceedings.

Richards, M. (2009). The national awareness and early diagnosis initiative in England: assembling the evidence. *British Journal of Cancer*, 101(S2):S1–S4.

Richardson, A. K. and Potter, J. D. (2014). Screening for colorectal cancer and prostate cancer: challenges for New Zealand. *The New Zealand Medical Journal (Online)*, 127(1395):23–30.

Robertson, C. and Boyle, P. (1998). Age–period–cohort analysis of chronic disease rates. I: modelling approach. *Statistics in Medicine*, 17(12):1305–1323.

Rosenbauer, J. and Strassburger, K. (2008). Comments on ?Age–period–cohort models for the Lexis diagram?by Carstensen B. Statistics in Medicine 2007; 26: 3018–3045. *Statistics in Medicine*, 27(9):1557–1561.

Rosenberg, P. S. (2019). A new age-period-cohort model for cancer surveillance research. *Statistical Methods in Medical Research*, 28(10-11):3363–3391.

Rosenberg, P. S. and Anderson, W. F. (2011). Age-period-cohort models in cancer surveillance research: ready for prime time? *Cancer Epidemiology Biomarkers & Prevention*, 20(7):1263–1268.

Rosenberg, P. S., Check, D. P., and Anderson, W. F. (2014). A web tool for age–period–cohort analysis of cancer incidence and mortality rates. *Cancer Epidemiology and Prevention Biomarkers*, 23(11):2296–2302.

Rutherford, M. J., Lambert, P. C., and Thompson, J. R. (2010). Age–period–cohort modeling. *The Stata Journal*, 10(4):606–627.

Sacerdote, C. and Ricceri, F. (2018). Epidemiological dimensions of the association between type 2 diabetes and cancer: A review of observational studies. *Diabetes Research and Clinical Practice*, 143(2018):369–377.

Sackett, D. L., Haynes, R. B., Tugwell, P., et al. (1985). *Clinical epidemiology: a basic science for clinical medicine.* Little, Brown and Company.

Salzmann, P., Kerlikowske, K., and Phillips, K. (1997). Cost-effectiveness of extending screening mammography guidelines to include women 40 to 49 years of age. *Annals of Internal Medicine*, 127(11):955–965.

Sammour, T., Kahokehr, A., Vather, R., Connolly, A., and Hill, A. (2010). Ethnic disparity in colonic cancer outcomes in New Zealand–biology or an access issue? *Colorectal Disease*, 12(7Online):e50–e56.

Samson, P., O'Grady, G., and Keating, J. (2009). An international comparison study of stage of colorectal cancer at diagnosis: how does New Zealand compare? *The New Zealand Medical Journal (Online)*, 122(1294).

Sandiford, P., Abdel-Rahman, M. E., Allemani, C., Coleman, M. P., and Gala, G. (2015). How many cancer deaths could New Zealand avoid if five-year relative survival ratios

were the same as in Australia? *Australian and New Zealand Journal of Public Health*, 39(2):157–161.

Sarfati, D., Garvey, G., Robson, B., Moore, S., Cunningham, R., Withrow, D., Griffiths, K., Caron, N. R., and Bray, F. (2018). Measuring cancer in indigenous populations. *Annals of Epidemiology*, 28(5):335–342.

Sarfati, D. and Jackson, C. (2020). Context of cancer control in New Zealand. *Journal of Cancer Policy*, 23(2020):100211.

Scott, N., Hill, J., Smith, J., Walker, K., Kuryba, A., van der Meulen, J., Greenaway, G., Yelland, A., and Meace, C. (2013). National Bowel Cancer Audit Annual Report 2013. `https://www.hqip.org.uk/wp-content/uploads/2018/02/national-bowel-cancer-audit-annual-report-2013.pdf`.

Secker, S. D., Atmore, C., Bramley, D., De Groot, C., Stevens, W., Sarfati, D., Brown, C., Hill, A., Reid, P., Lawrenson, R., et al. (2015). The PIPER Project. `https://www.fmhs.auckland.ac.nz/assets/fmhs/sms/ctnz/docs/THE%20PIPER%20PROJECT%20Final%20deliverable%20report%207%20August%202015%20(HRC%2011_764%20FINDLAY).pdf`.

Selvachandran, S., Hodder, R., Ballal, M., Jones, P., and Cade, D. (2002). Prediction of colorectal cancer by a patient consultation questionnaire and scoring system: a prospective study. *The Lancet*, 360(9329):278–283.

Shah, A. B., Sarfati, D., Blakely, T., Atkinson, J., and Dennett, E. R. (2012). Trends in colorectal cancer incidence rates in New Zealand, 1981–2004. *ANZ Journal of Surgery*, 82(4):258–264.

Sharples, K., Firth, M., Hinder, V., Hill, A., Jeffery, M., Sarfati, D., Brown, C., Atmore, C., Lawrenson, R., Reid, P., et al. (2018). The New Zealand PIPER Project: colorectal cancer survival according to rurality, ethnicity and socioeconomic deprivation?results from a retrospective cohort study. *The New Zealand Medical Journal*, 131(1476):24–39.

Shaw, C., Atkinson, J., and Blakely, T. (2009). (Mis) classification of ethnicity on the New Zealand Cancer Registry: 1981-2004. *The New Zealand Medical Journal (Online)*, 122(1294):10–22.

Sheerin, I., Green, T., Sarfati, D., and Cox, B. (2015). Projected costs of colorectal cancer treatment in New Zealand in the absence of population screening. *The New Zealand Medical Journal (Online)*, 128(1408):72–85.

Siegel, R., DeSantis, C., and Jemal, A. (2014). Colorectal cancer statistics, 2014. *CA: A Cancer Journal for Clinicians*, 64(2):104–117.

Siegel, R. L., Fedewa, S. A., Anderson, W. F., Miller, K. D., Ma, J., Rosenberg, P. S., and Jemal, A. (2017). Colorectal cancer incidence patterns in the United States, 1974–2013. *JNCI: Journal of the National Cancer Institute*, 109(8):1–6.

Siegel, R. L., Torre, L. A., Soerjomataram, I., Hayes, R. B., Bray, F., Weber, T. K., and Jemal, A. (2019). Global patterns and trends in colorectal cancer incidence in young adults. *Gut*, 68(12):2179–2185.

Sikdar, K. C., Walsh, S. J., Roche, M., Jiang, Y., Syrowatka, A., and Collins, K. D. (2013). Diabetes and sex-specific colorectal cancer risks in Newfoundland and Labrador: a population-based retrospective cohort study. *Canadian Journal of Public Health*, 104(2):101–107.

Simpson, G. (2018a). *gratia: Graceful 'ggplot'-Based Graphics and Other Functions for GAMs Fitted Using 'mgcv'*. R package version 0.3.1.

Simpson, G. L. (2018b). Modelling palaeoecological time series using generalised additive models. *Frontiers in Ecology and Evolution*, 6(2018):149.

Singh, H., Demers, A. A., Xue, L., Turner, D., and Bernstein, C. N. (2008). Time trends in colon cancer incidence and distribution and lower gastrointestinal endoscopy utilization in Manitoba. *American Journal of Gastroenterology*, 103(5):1249–1256.

Smith, G. C., Seaman, S. R., Wood, A. M., Royston, P., and White, I. R. (2014). Correcting for optimistic prediction in small data sets. *American Journal of Epidemiology*, 180(3):318–324.

Smith, T. R., Wakefield, J., et al. (2016). A review and comparison of age–period–cohort models for cancer incidence. *Statistical Science*, 31(4):591–610.

Sox, H., Blatt, M., Higgins, M., and Marton, K. (1988). Selection and interpretation of diagnostic tests. In *Medical Decision Making*, pages 239–290. Butterwoth-Heinemann Boston, Mass.

Stamm, R., Aluzaite, K., Arnold, M., Caspritz, T., White, C., and Schultz, M. (2020). Challenges for the future: the gastroenterology specialist workforce in New Zealand. *The New Zealand Medical Journal (Online)*, 133(1519):32–5.

Stanton, E. A. (2007). The human development index: A history. `https://scholarworks.umass.edu/cgi/viewcontent.cgi?article=1101&context=peri_workingpapers`.

Starup-Linde, J., Karlstad, O., Aistrup Eriksen, S., Vestergaard, P., K Bronsveld, H., de Vries, F., Andersen, M., Auvinen, A., Haukka, J., Hjellvik, V., et al. (2013). CARING (CAncer Risk and INsulin analoGues): the association of diabetes mellitus and cancer risk with focus on possible determinants-a systematic review and a meta-analysis. *Current Drug Safety*, 8(5):296–332.

Statistics NZ (2005). Understanding and Working with Ethnicity Data. `http://archive.stats.govt.nz/~/media/Statistics/browse-categories/population/census-counts/review-measurement-ethnicity/understanding-working-ethnicity-data.pdf`.

Statistics NZ (2018). Census 2018: Design of Forms. `https://www.stats.govt.nz/assets/Reports/2018-census-design-of-forms/2018-Census-Design-of-forms.pdf`.

Statistics NZ (2020a). Estimated Resident Population by Age and Sex (1991+) (Annual-Dec). `http://archive.stats.govt.nz/infoshare/`.

Statistics NZ (2020b). Sex and gender identity statistical standards: Consultation. `https://www.stats.govt.nz/consultations/sex-and-gender-identity-statistical-standards-consultation`.

Steine, S., Stordahl, A., Laerum, F., and Laerum, E. (1994). Referrals for double-contrast barium examination: factors influencing the probability of finding polyps or cancer. *Scandinavian Journal of Gastroenterology*, 29(3):260–264.

Steyerberg, E. W. et al. (2019). *Clinical prediction models*. Springer.

Steyerberg, E. W. and Vergouwe, Y. (2014). Towards better clinical prediction models: seven steps for development and an ABCD for validation. *European Heart Journal*, 35(29):1925–1931.

Suissa, S. and Azoulay, L. (2012). Metformin and the risk of cancer: time-related biases in observational studies. *Diabetes Care*, 35(12):2665–2673.

Sundborn, G., Metcalf, P., Scragg, R., Schaaf, D., Dyall, L., Gentles, D., Black, P., and Jackson, R. (2007). Ethnic differences in the prevalence of new and known diabetes mellitus, impaired glucose tolerance and impaired fasting glucose. Diabetes heart and health survey (DHAH) 2002-2003, Auckland New Zealand. *The New Zealand Medical Journal (Online)*, 120(1257):1–12.

Sung, H., Siegel, R. L., Rosenberg, P. S., and Jemal, A. (2019). Emerging cancer trends among young adults in the USA: analysis of a population-based cancer registry. *The Lancet Public Health*, 4(3):e137–e147.

Svensson, E., Grotmol, T., Hoff, G., Langmark, F., Norstein, J., and Tretli, S. (2002). Trends in colorectal cancer incidence in Norway by gender and anatomic site: an age–period–cohort analysis. *European Journal of Cancer Prevention*, 11(5):489–495.

Svensson, E., Møller, B., Tretli, S., Barlow, L., Engholm, G., Pukkala, E., Rahu, M., and Tryggvadottir, L. (2005). Early life events and later risk of colorectal cancer: age-period-cohort modelling in the Nordic countries and Estonia. *Cancer Causes & Control*, 16(3):215–223.

Swart, E. M., Sarfati, D., Cunningham, R., Dennett, E., Signal, V., Gurney, J., and Stanley, J. (2013). Ethnicity and rectal cancer management in New Zealand. *The New Zealand Medical Journal (Online)*, 126(1384):42–52.

Szklo, M. (1998). Population-based cohort studies. *Epidemiologic Reviews*, 20(1):81–90.

Teng, A. M., Atkinson, J., Disney, G., Wilson, N., Sarfati, D., McLeod, M., and Blakely, T. (2016). Ethnic inequalities in cancer incidence and mortality: census-linked cohort studies with 87 million years of person-time follow-up. *BMC Cancer*, 16(1):1–14.

Therneau, T. M., Atkinson, E. J., et al. (1997). An introduction to recursive partitioning using the RPART routines. Technical report, Mayo Foundation.

Thompson, M. (2002). ACPGBI Referral guidelines for colorectal cancer. *Colorectal Disease*, 4(4):287–297.

Thompson, M., Flashman, K., Wooldrage, K., Rogers, P., Senapati, A., O'leary, D., and Atkin, W. (2008). Flexible sigmoidoscopy and whole colonic imaging in the diagnosis of cancer in patients with colorectal symptoms. *British Journal of Surgery*, 95(9):1140–1146.

Thörn, M., Bergström, R., Kressner, U., Sparén, P., Zack, M., and Ekbom, A. (1998). Trends in colorectal cancer incidence in Sweden 1959-93 by gender, localization, time period, and birth cohort. *Cancer Causes & Control*, 9(2):145–152.

Thorne, K., Hutchings, H. A., and Elwyn, G. (2006). The effects of the Two-Week Rule on NHS colorectal cancer diagnostic services: a systematic literature review. *BMC Health Services Research*, 6(1):43.

Torre, L. A., Siegel, R. L., Ward, E. M., and Jemal, A. (2016). Global cancer incidence and mortality rates and trends?an update. *Cancer Epidemiology and Prevention Biomarkers*, 25(1):16–27.

Tsilidis, K. K., Kasimis, J. C., Lopez, D. S., Ntzani, E. E., and Ioannidis, J. P. (2015). Type 2 diabetes and cancer: umbrella review of meta-analyses of observational studies. *British Medical Journal*, 350(2015):g7607.

Turner, R. M., Cameron, C., and Samaranayaka, A. (2019). Understanding receiver operator characteristic (ROC) curves. *New Zealand Medical Student Journal*, 29(29):50–51.

Vader, J.-P., Froehlich, F., Dubois, R., Beglinger, C., Wietlisbach, V., Pittet, V., Ebel, N., Gonvers, J.-J., and Burnand, B. (1999). European Panel on the Appropriateness of Gastrointestinal Endoscopy (EPAGE): conclusion and WWW site. *Endoscopy*, 31(08):687–694.

Valent, F. (2015). Diabetes mellitus and cancer of the digestive organs: an Italian population-based cohort study. *Journal of Diabetes and its Complications*, 29(8):1056–1061.

Vandenbroucke, J. P., von Elm, E., Altman, D. G., Gøtzsche, P. C., Mulrow, C. D., Pocock, S. J., Poole, C., Schlesselman, J. J., Egger, M., Initiative, S., et al. (2014). Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): explanation and elaboration. *International Journal of Surgery*, 12(12):1500–1524.

VanderWeele, T. J. (2009). On the distinction between interaction and effect modification. *Epidemiology*, 20(6):863–871.

Vega, P., Valentín, F., and Cubiella, J. (2015). Colorectal cancer diagnosis: pitfalls and opportunities. *World Journal of Gastrointestinal Oncology*, 7(12):422–33.

Wahba, G. (1990). *Spline models for observational data*, volume 59. Siam.

Walter, F., Webster, A., Scott, S., and Emery, J. (2012). The Andersen Model of Total Patient Delay: a systematic review of its application in cancer diagnosis. *Journal of Health Services Research & Policy*, 17(2):110–118.

Wang, M., Hu, R.-Y., Wu, H.-B., Pan, J., Gong, W.-W., Guo, L.-H., Zhong, J.-M., Fei, F.-R., and Yu, M. (2015). Cancer risk among patients with type 2 diabetes mellitus: a population-based prospective study in China. *Scientific Reports*, 5(1):1–7.

Wheeler, B. J., Braund, R., Galland, B., Mikuscheva, A., Wiltshire, E., Jefferies, C., and de Lange, M. (2019). District health board of residence, ethnicity and socioeconomic status all impact publicly funded insulin pump uptake in New Zealand patients with type 1 diabetes. *The New Zealand Medical Journal*, 132(1491):78–89.

Williams, T. G., Cubiella, J., Griffin, S. J., Walter, F. M., and Usher-Smith, J. A. (2016). Risk prediction models for colorectal cancer in people with symptoms: a systematic review. *BMC Gastroenterology*, 16(1):1–16.

Windner, Z., Crengle, S., de Graaf, B., Samaranayaka, A., and Derrett, S. (2018). New Zealanders' experiences and pathways to a diagnosis of bowel cancer: a cross-sectional descriptive study of a younger cohort. *The New Zealand Medical Journal*, 131(1483):30–39.

Wong, M. C., Huang, J., Lok, V., Wang, J., Fung, F., Ding, H., and Zheng, Z.-J. (2020). Differences in Incidence and Mortality Trends of Colorectal Cancer, Worldwide, Based on Sex, Age, and Anatomic Location. *Clinical Gastroenterology and Hepatology*.

Wood, G. (2002). Assessing goodness of fit for Poisson and negative binomial models with low mean. *Communications in Statistics: Theory and Methods*, 31(11):1977–2001.

Wood, S. N. (2003). Thin plate regression splines. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(1):95–114.

Wood, S. N. (2017). *Generalized additive models: an introduction with R*. CRC press.

World Health Organization (2021). Gender and health. `https://www.who.int/health-topics/gender#tab=tab_1`.

Wu, L., Yu, C., Jiang, H., Tang, J., Huang, H.-l., Gao, J., and Zhang, X. (2013). Diabetes mellitus and the occurrence of colorectal cancer: an updated meta-analysis of cohort studies. *Diabetes Technology & Therapeutics*, 15(5):419–427.

Xu, H., Stenner, S. P., Doan, S., Johnson, K. B., Waitman, L. R., and Denny, J. C. (2010). MedEx: a medication information extraction system for clinical narratives. *Journal of the American Medical Informatics Association*, 17(1):19–24.

Yang, Y. and Land, K. C. (2013). *Age-period-cohort analysis: New models, methods, and empirical applications*. CRC Press.

Yang, Y.-X., Hennessy, S., and Lewis, J. D. (2004). Insulin therapy and colorectal cancer risk among type 2 diabetes mellitus patients. *Gastroenterology*, 127(4):1044–1050.

Yeoman, A. and Parry, S. (2010). A survey of colonoscopy capacity in New Zealand?s public hospitals. *The New Zealand Medical Journal*, 120(1258):12–20.

Yin, S., Bai, H., and Jing, D. (2014). Insulin therapy and colorectal cancer risk among type 2 diabetes mellitus patients: a systemic review and meta-analysis. *Diagnostic Pathology*, 9(1):1–6.

Zauber, A. G. (2015). The impact of screening on colorectal cancer mortality and incidence: has it really made a difference? *Digestive Diseases and Sciences*, 60(3):681–691.

Appendices

Appendix A

Ethics approval

Health and Disability Ethics Committees
Ministry of Health
133 Molesworth Street
PO Box 5013
Wellington
6011

0800 4 Ethics
hdecs@moh.govt.nz

16 July 2018

Ms Malgorzata Hirsz
University of Waikato
Hillcrest Road
Hamilton 3240

Dear Ms Hirsz

Re:	**Ethics ref:**	**18/CEN/118**
	Study title:	Investigation of the risk of colorectal cancer in patients with type 2 diabetes in relation to insulin use and possible confounders

I am pleased to advise that this application has been *approved* by the Central Health and Disability Ethics Committee. This decision was made through the HDEC-Expedited Review pathway.

Conditions of HDEC approval

HDEC approval for this study is subject to the following conditions being met prior to the commencement of the study in New Zealand. It is your responsibility, and that of the study's sponsor, to ensure that these conditions are met. No further review by the Central Health and Disability Ethics Committee is required.

Standard conditions:

1. Before the study commences at *any* locality in New Zealand, all relevant regulatory approvals must be obtained.

2. Before the study commences at *each given* locality in New Zealand, it must be authorised by that locality in Online Forms. Locality authorisation confirms that the locality is suitable for the safe and effective conduct of the study, and that local research governance issues have been addressed.

After HDEC review

Please refer to the *Standard Operating Procedures for Health and Disability Ethics Committees* (available on www.ethics.health.govt.nz) for HDEC requirements relating to amendments and other post-approval processes.

Your next progress report is due by 15 July 2019.

Participant access to ACC

The Central Health and Disability Ethics Committee is satisfied that your study is not a clinical trial that is to be conducted principally for the benefit of the manufacturer or distributor of the medicine or item being trialled. Participants injured as a result of treatment received as part of your study may therefore be eligible for publicly-funded compensation through the Accident Compensation Corporation (ACC).

Please don't hesitate to contact the HDEC secretariat for further information. We wish you all the best for your study.

Yours sincerely,

Mrs Helen Walker
Chairperson
Central Health and Disability Ethics Committee

Encl: appendix A: documents submitted
appendix B: statement of compliance and list of members

Appendix B

Ethics approval (amendment)

Health and Disability Ethics Committees

Health and Disability Ethics Committees
Ministry of Health
133 Molesworth Street
PO Box 5013
Wellington
6011

0800 4 ETHICS
hdecs@moh.govt.nz

13 September 2018

Ms Malgorzata Hirsz
University of Waikato
Hillcrest Road
Hamilton 3240

Dear Ms Hirsz,

Re:	**Ethics ref:**	**18/CEN/118/AM01**
	Study title:	Investigation of the risk of colorectal cancer in patients with type 2 diabetes in relation to insulin use and possible confounders

I am pleased to advise that this amendment has been *approved* by the Central Health and Disability Ethics Committee. This decision was made through the HDEC Expedited Review pathway.

Please don't hesitate to contact the HDEC secretariat for further information. We wish you all the best for your study.

Yours sincerely,

Mrs Helen Walker
Chairperson
Central Health and Disability Ethics Committee

Encl: appendix A: documents submitted
appendix B: statement of compliance and list of members

Appendix A
Documents submitted and approved

Document	Version	Date
Protocol: This is the corrected protocol. The changes are only in the methods section.	2	27 August 2018
Post Approval Form	AM01	27 August 2018

Appendix B
Statement of compliance and list of members

Statement of compliance

The Central Health and Disability Ethics Committee:

- is constituted in accordance with its Terms of Reference
- operates in accordance with the *Standard Operating Procedures for Health and Disability Ethics Committees*, and with the principles of international good clinical practice (GCP)
- is approved by the Health Research Council of New Zealand's Ethics Committee for the purposes of section 25(1)(c) of the Health Research Council Act 1990
- is registered (number 00008712) with the US Department of Health and Human Services' Office for Human Research Protection (OHRP).

List of members

Name	*Category*	*Appointed*	*Term Expires*
Mrs Helen Walker	Lay (consumer/community perspectives)	01/07/2015	01/07/2018
Dr Peter Gallagher	Non-lay (health/disability service provision)	30/07/2015	30/07/2018
Mrs Sandy Gill	Lay (consumer/community perspectives)	30/07/2015	30/07/2018
Dr Patries Herst	Non-lay (intervention studies)	27/10/2015	27/10/2018
Dr Dean Quinn	Non-lay (intervention studies)	27/10/2015	27/10/2018
Dr Cordelia Thomas	Lay (the law)	20/05/2017	20/05/2020

Unless members resign, vacate or are removed from their office, every member of HDEC shall continue in office until their successor comes into office (HDEC Terms of Reference)

http://www.ethics.health.govt.nz

Appendix C

Ratification letter

22 March 2021

Malgorzata Hirsz
By email: mh331@students.waikato.ac.nz

Dear Malgorzata

HDEC Ethics 18/CEN/118 Investigation of the risk of colorectal cancer in patients with type 2 diabetes in relation to insulin use and possible confounders

Thank you for submitting your HDEC approval for your project 18/CEN/118 including the amendment for analysis of data from Cancer Register from 1994-2018 requested in October 2019.

We are pleased to ratify this approval, as per Dr Julie Barbour's email to you dated 30 October 2019.

Regards,

Emeritus Professor Roger Moltzen MNZM
Chairperson
University of Waikato Human Research Ethics Committee

Appendix D

Correspondence with MoH (ethics for sub-study 1)

This appendix contains my correspondence with MoH (via email) in order to get approval for addressing the objectives included in sub-study 1 which led to the Ratification letter in Appendix C.

Date: Sun, 6 Oct 2019 10:56:44
Subject: Question about amendment of ethics application
From: Malgorzata Hirsz <mh331@students.waikato.ac.nz>
To: hdecs@health.govt.nz

Kia ora,

My name is Malgorzata Hirsz. I am a PhD student at the University of Waikato. I have approval for the study which I am conducting, number 18/CEN/118/AM01.

I would like to ask if I need to amend my application in order to add one objective to the study. The study objectives are listed on page 11 of the protocol. I want to address the following additional objective:

To investigate if there are any demographic groups within the New Zealand population which were in particularly high risk of being diagnosed with colorectal cancer (with relation to the stage at diagnosis) within the period 1994-2018 and how it has changed during this period.

The data required for this analysis are colorectal cancer registrations from NZ Cancer Registry for the years 1994-2018. I have already permission from HDEC for analysis of the NZCR data for the years 1994-2018 for addressing other objectives of my study, and therefore I do not need any additional data.

There is no additional risk for compromising patients' privacy beyond what was already declared in the main ethics application.

Could you, please, advise me if I need to amend my ethics application?

Many thanks
Malgorzata Hirsz

From: hdecs@health.govt.nz
To: Malgorzata Hirsz <mh331@students.waikato.ac.nz>
Subject: Re: Question about amendment of ethics application
Sender: Mark.Joyce@health.govt.nz
Date: Mon, 7 Oct 2019 16:09:20

Hi Malgorzata,

Thanks for your question. We would consider this additional analysis a significant change to your study design, and should therefore be submitted

as an amendment. A potential ethical issue to consider will be the risk of stigmatisation during the reporting of your study if results indicate higher risk for some demographics.

Kind regards,

Mark

Mark Joyce
Advisor
Ethics
Quality Assurance and Safety
Health System Improvement and Innovation
Ministry of Health

http://www.health.govt.nz
mailto:Mark.Joyce@health.govt.nz

Hi Mark,

Thank you for your email and for the explanation.
I will think and talk to my supervisors about the risk of stigmatisation during the reporting of the study results.

Kind regards
Malgorzata

Date: Mon, 28 Oct 2019 23:07:20
Subject: Re: Question about amendment of ethics application
From: Malgorzata Hirsz <mh331@students.waikato.ac.nz>
To: hdecs@health.govt.nz

Hi Mark,

I am writing again to ask for your assistance in relation to my query from 6th October.

After your advice I talked to my supervisor and I was advised to submit a new ethics application. In my supervisor's opinion an amendment to my existing project would not be sufficient.

I filled in the form for the second part of my study and I found out that HDEC approval is not required because the Ministry of Health, before providing the data to me, encrypted patients' NHI numbers, and therefore I cannot identify any individual.

My application is ready, and the protocol is nearly updated, however I do not want to ask for peer review if it is not required as the reviewer has to spend time unnecessarily on my application.

Could you tell me, please, if I could talk to you over the phone to discuss the issue or alternatively, can you advise me what to do (e.g. should I talk to the University Ethics Committee) ?

Kind regards
Malgorzata

Sender: Mark.Joyce@health.govt.nz
Date: Wed, 30 Oct 2019 16:02:31

Hi again Malgorzata,

I've just had a quick look through your original application, and it might be the case that this did not necessarily require HDEC review either (of course, the advisor may have made a judgement call based on the linking of multiple datasets, which does raise the risk of identification). Like you say, MoH provides data with encrypted NHIs only, and if you will be unable to identify individuals this reduces the risk threshold below that which necessitates submission to HDEC.

I'm just about to leave for the day, but you can certainly call me on 04 816 2351 on Friday (I'm away tomorrow). I would however advise going through your institutional ethics committee, and if they insist on an out-of-scope letter from HDEC we can certainly provide this.

All the best,

Mark

Mark Joyce
Advisor
Ethics
Quality Assurance and Safety
Health System Improvement and Innovation
Ministry of Health

Appendix E

Māori Consultation

Te Puna Oranga Māori Consultation Research Review Committee

2 July 2018

Re: Māori Consultation for 'How can we improve time to diagnosis of colorectal cancer in symptomatic patients? Symptoms and patient characteristics as criteria for diagnosis of colorectal cancer in primary and secondary care in New Zealand.'

Name of Applicant: Malgorzata Hirsz

Tēnā Koe Malgorzata,

Thank you for submitting the above research proposal to the Waikato DHB Te Puna Oranga Māori Health Research Committee for Māori consultation. The research application has been reviewed in order to support and prompt the researcher to think about how this research will improve health outcomes and eliminate inequity for Māori living within the Waikato DHB region.

1. The Committee acknowledges the researchers for collecting ethnicity data as part of a demographic background of the participant to improve data collection for Māori in order to improve Māori health outcomes and reduce inequity for Māori.

2. The Committee encourages the research team to actively recruit equal numbers of Māori and Non-Māori. Any Research that involves Māori participation would require sufficient face to face time for fully informed consent to occur. Inclusion of the whānau of the Māori participant should be encouraged to support the continued engagement of the Maori participant in the research process.

3. The Committee encourages all research that involves participation of individuals, especially Māori participants to fully inform them regarding the detail of tissue collection. One consent form for the current use of Tissue. One consent form for the future use of tissue (this should be clear to the participant).

4. If cultural issues arise for the Māori participant during any research, they will inform the research team during the study that an issue has occurred. Cultural issues may not be obvious to the participant or the researcher prior to commencement of the research.

5. The Committee encourages the research team to continue to consult with Te Puna Oranga, Māori Health service at any time, should they have any further queries.

6. Feedback regarding this research is appreciated and can be shared back to the Kaunihera Kaumatua via Te Puna Oranga Māori Health Service

The Committee endorses this research proposal with the consideration of the above cultural recommendations where appropriate and encourages the researcher to collect ethnicity data for all study participants seen at Waikato DHB for our own internal records.

The Committee suggested that you should seek a Maori epidemiology supervisor for your study, and Dr Nina Scott has offered to help you to do this.

Millie Berryman
Kaitakawaenga Māori
Te Puna Oranga-Maori Health Service
Millie.Berryman@waikatodhb.health.nz

Appendix F

Example of the use of predicted IRs from APC model

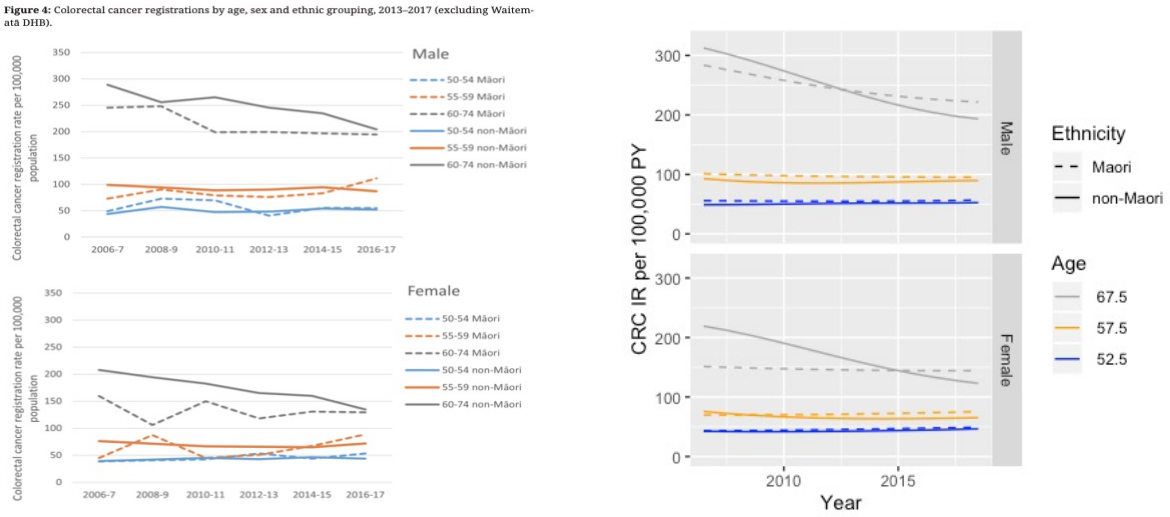


Figure F.1: *Left figure: age-specific rates from McLeod et al. (2021) for three age brackets, right figure: incidence rates based on an age-period-cohort model fitted to CRC incidence data 2006–2018 presented for the mid-point of the three age brackets.*